THINKING BEYOND VIDEOS: UNIFYING VIDEO REASONING AND DEEP RESEARCH FOR OPEN-WORLD VIDEO AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-world video understanding often requires locating sparse visual evidence in a video and acquiring knowledge that is contained neither in the video itself nor in the model's parameters. Thinking-with-Videos actively localizes such evidence, and Deep Research acquires such knowledge over multiple search steps, but the two directions are developed in isolation. To bridge this gap, we introduce `VideoRover`, a unified Video Deep Research framework that iteratively coordinates video cropping, multimodal search, and webpage browsing. Given a video–question pair, `VideoRover` conditions each action on the observations accumulated so far, so that localized clips anchor external search, while retrieved evidence triggers further video inspection and verification. Existing data rarely cover such trajectories, so we build an automated pipeline that constructs questions jointly dependent on video and web evidence and synthesizes verified interaction trajectories, yielding 26K SFT examples and 3K hard RL instances. We also introduce `VideoRover-Bench`, a benchmark stratified by video duration and research difficulty. Experiments on VideoDR and `VideoRover-Bench` show `VideoRover`-8B-RL performs on par with proprietary models in direct-answer settings while outperforming some larger models given the same tools. Additional experiments further demonstrate broad generalization and adaptive coordination between video grounding and external retrieval.

1 INTRODUCTION

Real-world multimodal tasks often require both fine-grained visual evidence and open-world knowledge, making them difficult to solve from the fixed inputs and parametric knowledge of multimodal large language models (MLLMs) (Abootorabi et al., 2025; Xu & Peng, 2025; Shi et al., 2025). Tool-augmented agents address these limitations along two complementary directions. For *active perception*, Thinking-with-Images (Zheng et al., 2026; Zhang et al., 2026c; Shen et al., 2026) lets models inspect visual regions during reasoning, and Thinking-with-Videos (Meng et al., 2026; Zhang et al., 2026b) extends this capability to temporal localization, segment cropping, and adaptive resampling. For *information acquisition*, retrieval-augmented generation (RAG) (Chen et al., 2022; Wang et al., 2024) supplies external knowledge, and search agents (Jin et al., 2025; Wu et al., 2026) and Deep Research (Zheng et al., 2025b; Li et al., 2026b) extend it with multi-round text and image search, webpage browsing, and evidence verification. *Active perception* decides what evidence to observe in the input, whereas *information acquisition* decides what knowledge to seek beyond it.

Open-world video tasks require the two to operate jointly. In news, documentaries, tutorials, and lectures, briefly appearing entities or clues distributed across distant segments must be linked to background, historical, or up-to-date information (Liang et al., 2025; Liu et al., 2026a; Gao et al., 2026; Fang et al., 2026). Thinking-with-Videos can locate relevant visual cues, but the information needed to answer the question may be absent from the video. External retrieval can fill this gap, yet effective search depends on grounding the query in the correct visual anchor. However, when the two are directly combined in a fixed-order pipeline of video sampling, query generation, and web search, early perception errors propagate into retrieval, while retrieved evidence cannot guide further inspection of uncertain video content. We argue that *Video Deep Research* should instead allow video observations to guide retrieval and retrieved evidence to refine subsequent video inspection.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

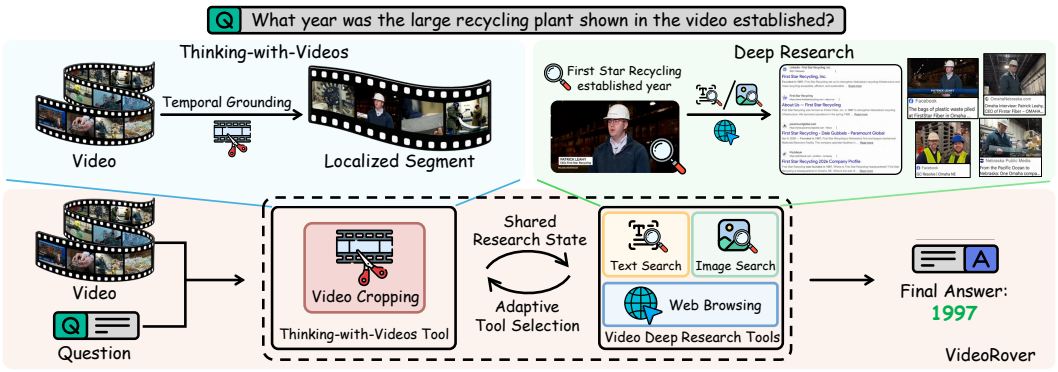


Figure 1: Illustration of how `VideoRover` unifies Thinking-with-Videos and Deep Research for open-world video questions. While the first two separately localize video evidence or retrieve external knowledge, `VideoRover` maintains a shared research state to adaptively coordinate video cropping, text and image search, and webpage browsing toward the final answer.

Despite progress along both directions, three challenges remain. **First**, temporal grounding must convert sparse evidence in long videos into precise anchors for external retrieval. Existing Thinking-with-Videos methods localize segments, select keyframes, and resample adaptively (Zhang et al., 2026a; Zeng et al., 2026; Liu et al., 2026b), but use these observations only for in-video reasoning, never turning them into external queries. **Second**, research planning must choose appropriate retrieval actions and construct queries from cues available only after grounding. Existing Deep Research agents operate mainly over text, static images, and webpages (Jin et al., 2025; Wu et al., 2026; Huang et al., 2026), so neither tool selection nor query construction is guided by key visual evidence localized within video frames. **Third**, video and retrieval actions must write to a shared state, so that a retrieval result can send the model back to the video and the resulting observation can guide the next query. Existing methods do not support this interaction because they treat video reasoning and retrieval as separate processes without iterative feedback, while existing datasets rarely provide complete trajectories covering grounding, retrieval, evidence revision, and answer synthesis, thereby limiting supervision for learning such coordination. Therefore, attaching search tools to a video reasoner does not by itself produce the coordination that Video Deep Research requires.

To address these challenges, we introduce `VideoRover`, a unified Video Deep Research framework for joint video grounding and multimodal retrieval (Figure 1). For reliable temporal grounding, `VideoRover` first localizes and densely inspects a segment, turning sparse evidence into a visual anchor. For adaptive research planning, it selects a keyframe for image retrieval, constructs text queries for missing knowledge, and visits webpages according to the current *evidence gap*. For iterative coordination, every tool result updates a shared research state and determines whether to continue searching, re-localize or re-examine video evidence, or answer. Video observations thereby ground external search, while retrieved evidence can redirect video reasoning. Considering no existing corpus records such trajectories, we develop an automated pipeline that constructs questions jointly dependent on video and external knowledge and synthesizes verified interaction trajectories, yielding 26K SFT examples and 3K hard RL instances. We further introduce `VideoRover-Bench`, stratified by video duration and task difficulty, to evaluate grounding, retrieval, and multi-source reasoning across different temporal and research complexities.

Our main contributions are summarized as follows:

- We propose `VideoRover`, which unifies Thinking-with-Videos and Deep Research in an agent that uses each tool result to coordinate video observation and open-world retrieval.
- We develop an automated pipeline for constructing Video Deep Research questions and verified trajectories involving temporal localization, multimodal retrieval, and evidence synthesis.
- We introduce `VideoRover-Bench`, a stratified benchmark spanning different video durations, task difficulty levels, and reasoning complexities.
- Experiments on VideoDR, `VideoRover-Bench`, and general video benchmarks demonstrate strong performance, while ablation and tool-use analyses validate the complementary tool design and adaptive coordination between video inspection and external retrieval.

2 Related Work

Thinking-with-Videos. Recent multimodal large language models (MLLMs) have advanced video understanding and reasoning, yet uniform sampling under a fixed visual token budget can easily miss sparse evidence in long videos. LongVA (Zhang et al., 2024) and LongVILA (Chen et al., 2025) increase context capacity, while token compression and hierarchical modeling improve efficiency, but these approaches typically determine the visual input before inference and cannot adapt observation to intermediate reasoning needs. Tool-augmented methods instead treat visual content as an active reasoning workspace: OpenThinkIMG (Su et al., 2025), DeepEyes (Zheng et al., 2026; Hong et al., 2026), and Thyme (Zhang et al., 2026c) enable iterative image inspection, while VITAL (Zhang et al., 2026a), LongVT (Yang et al., 2026), FrameThinker (He et al., 2026), Video-o3 (Zeng et al., 2026) and VideoTemp-o3 (Liu et al., 2026b) extend this idea to temporal localization, keyframe selection, and adaptive video resampling. Despite improving access to sparse temporal evidence, existing Thinking-with-Videos methods primarily reason within the input video. `VideoRover` further connects localized video evidence to image search and external retrieval, allowing retrieved knowledge to verify and redirect subsequent video reasoning.

Deep Research. Early search agents such as WebGPT (Nakano et al., 2021) and Search-R1 (Jin et al., 2025) extend retrieval-augmented generation with iterative reasoning and search, but remain primarily text-centric. Recent multimodal systems further incorporate visual search and interaction: MMSearch-R1 (Wu et al., 2026) and WebWatcher (Geng et al., 2026) learn multimodal search behaviors through reinforcement learning, while Vision-DeepResearch (Huang et al., 2026), MM-DeepResearch (Yao et al., 2026), OpenSearch-VL (Chen et al., 2026), and HyperEyes (Li et al., 2026a) further support long-horizon, multi-tool visual–textual exploration with capabilities such as multi-entity search, visual manipulation, and cross-modal evidence synthesis. These methods substantially extend Deep Research beyond text, yet their perceptual inputs remain largely static images or webpages, without explicitly addressing sparse evidence distributed over long videos. `VideoRover` targets this gap by placing temporal localization, image search, text search, and webpage browsing in a unified trajectory, where video evidence grounds external retrieval and retrieved knowledge can trigger video revisiting and verification.

3 Task Formulation

A Video Deep Research instance is $x = (V, Q)$, where $V = (v_1, \ldots, v_N) \in \mathcal{V}$ is an N-frame video with v_i as the i-th frame, and $Q \in \mathcal{Q}$ is a research question. Given an open-world web environment $\mathcal{W}$, the goal is to produce a final answer $Y \in \mathcal{Y}$ by identifying the relevant video evidence and acquiring missing external knowledge. At step t, we represent the current research state as:

$$S_t = (G_t^v, G_t^w, h_t), \tag{1}$$

where G_t^v and G_t^w are the accumulated video and web evidence, and h_t is the interaction history. Given Q and S_t, the model samples the next action u_{t+1} from policy π_θ to address the remaining evidence gap, then updates the state with tool observation z_{t+1}:

$$u_{t+1} \sim \pi_\theta(\cdot \mid Q, S_t), \qquad S_{t+1} = \mathrm{Update}(S_t, u_{t+1}, z_{t+1}). \tag{2}$$

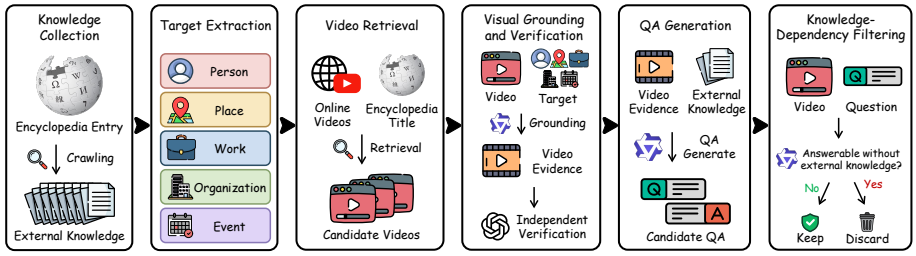
`VideoRover` uses four tools. The $\mathrm{crop_video}(V, [s_t, e_t])$ tool returns a densely sampled clip between start and end timestamps s_t and e_t. The $\mathrm{image_search}(v_{k_t})$ tool retrieves visually related information using the selected keyframe v_{k_t}, while $\mathrm{text_search}(q_t)$ takes a text query q_t and returns candidate webpages and summaries. Finally, $\mathrm{visit}(l_t)$ opens the webpage URL l_t to collect detailed evidence. See Appendix B for details. These tools may be invoked repeatedly until the model executes the terminal action answer. We denote the complete interaction trajectory as:

$$\tau = \big(S_0, u_1, z_1, S_1, \ldots, u_T, z_T, S_T, Y\big), \tag{3}$$

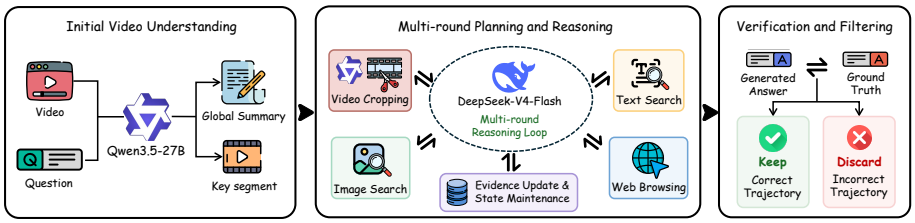
where S_0 is the initial research state containing the input $x = (V, Q)$ and a sparse sample of V, and T is the number of interaction steps. The process terminates when the accumulated evidence supports an answer Y or the maximum interaction budget is reached.

4 Data Synthesis and Benchmark Construction

High-quality data is essential for developing Video Deep Research capabilities. Unlike conventional video question answering, Video Deep Research requires not only reasoning over video content, but



(a) Question construction.



(b) Trajectory construction.

Figure 2: Pipelines for constructing Video Deep Research questions and tool-use trajectories.

also acquiring external information beyond the model's parametric knowledge and connecting video evidence with web evidence through multi-step interactions. To support this setting, we develop a data pipeline covering Video Deep Research question construction, SFT trajectory synthesis, and hard-example selection for RL. We further construct a stratified Video Deep Research benchmark, `VideoRover-Bench`, organized by video duration and task difficulty.

4.1 Question Construction

Figure 2a summarizes how we construct instances $x = (V, Q)$ with reference answers Y, following Section 3. Each instance enforces two dependencies: the video identifies a target referenced only indirectly by the question, while external web knowledge provides the missing information needed for the answer. Thus, neither video-only reasoning nor question-only web search is sufficient.

Knowledge collection and target extraction. We collect Wikipedia[1] articles and extract visually identifiable people, places, organizations, works, products, and events as candidate targets. For each target, we retain the article title, supporting text, and source URL as verifiable evidence for question generation and answer checking, while discarding candidates without clear factual answers or reliable sources, thereby providing reliable knowledge anchors for subsequent stages.

Video retrieval and visual grounding. We query YouTube[2] with each target and its Wikipedia title and retain the top 3 candidate videos. To remove videos that match the target only in their metadata, Qwen3.5-27B (Qwen Team, 2026a) analyzes 512 densely sampled frames and subtitles to localize likely intervals and keyframes. GPT-5.4-mini (Singh et al., 2025) then independently verifies the selected keyframe against the target identity. Only video–target pairs passing both stages are retained as reliable visual anchors for subsequent question answering.

Video-grounded question generation. Given a grounded video–target pair and collected web evidence, we generate Q by referring to the target indirectly (e.g., "a book cover displayed in the video") and asking about background, history, or recent developments. The identity is withheld so that answering the question requires grounding the target in V. We then disable retrieval and ask Qwen3-VL-4B (Bai et al., 2025a) to answer using only V and Q, removing samples it answers correctly. This filtering ensures that the final answer depends on both video evidence and external retrieval.

4.2 Trajectory Generation

High-quality interaction trajectories are essential for teaching the model Video Deep Research behaviors rather than only final-answer generation. Figure 2b summarizes our trajectory construction

[1] https://www.wikipedia.org/
[2] https://www.youtube.com/

process, which coordinates video observation, research planning, and evidence-guided tool use before filtering trajectories by answer correctness. We detail each stage below.

Roles and initialization. Each trajectory is jointly generated by DeepSeek-V4-Flash[3] (Xu et al., 2026) as the research planner and Qwen3.5-27B (Qwen Team, 2026a) as the video observer. Because DeepSeek-V4-Flash is a text-only model without visual perception, the observer inspects sampled frames and cropped segments and provides temporally grounded descriptions to the planner. We use the same observer as in question construction to maintain consistent visual grounding. During initialization, the observer summarizes sparsely sampled frames and proposes a relevant interval. Based on this summary, the planner's first response must localize a candidate key segment and invoke `crop_video`. The observer then examines the densely sampled crop for relevant entities, actions, text, and temporal cues. This ensures that subsequent retrieval is grounded in fine-grained video evidence rather than the initial sparse overview.

Evidence-guided tool use and filtering. After each crop, the planner selects the next action based on the visual observations and remaining evidence gap. If the target is not located, it localizes and crops another segment. When the crop reveals a visual entity, the planner prioritizes `image_search` to verify its identity and reduce visual misidentification. Once the identity is established, it invokes `text_search` if the required background, historical, or up-to-date information remains missing. When either search tool returns a relevant result, the planner calls `visit` to inspect the source webpage and extract detailed evidence rather than rely on search-result summaries. Each tool result updates the research state, enabling the planner to confirm, reject, or revise existing hypotheses and select its next action. The interaction ends when video and external evidence jointly support an answer. We retain only trajectories whose final answers match the ground truth.

4.3 Training Data

SFT Data. The synthesis process independently produces answers consistent with the ground truth, providing an additional check of question validity and trajectory correctness. We partition the verified trajectories by interaction length and retain those with at most 10 tool calls for supervised fine-tuning, yielding 26K SFT trajectories. These data cover the fundamental Video Deep Research behaviors of video grounding, multimodal retrieval, evidence updating, and answer synthesis, while longer trajectories are reserved as candidates for subsequent stages.

RL Data. After the cold start of SFT, we sample 5 complete rollouts per candidate question whose reference trajectory requires more than 10 tool calls, recording answer correctness, valid tool usage, and interaction length. We remove relatively easy questions answered correctly in more than 3 rollouts, then balance the remaining hard examples by visual-grounding difficulty, search frequency, and overall interaction complexity. This produces 3K RL instances whose rollout groups typically contain both successful and unsuccessful behaviors, providing informative within-group reward variation for relative policy optimization and emphasizing long-horizon autonomous decision making.

4.4 Benchmark Construction

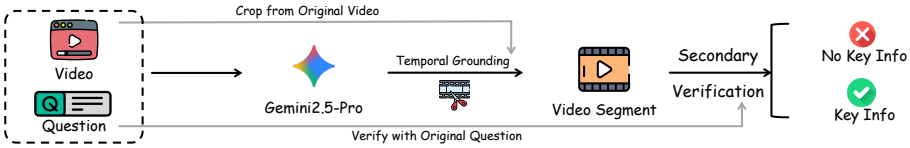
We construct `VideoRover-Bench` to evaluate Video Deep Research across temporal scales and research complexities. Figure 3 outlines verification, stratification, and balanced sampling.

Starting from the candidate video–question pool, we use Gemini-2.5-Pro (Comanici et al., 2025), which is independent of the data-generation models, to check question clarity and localize the relevant video segment. The localized segment is then paired with the question for a subsequent verification pass that determines whether it contains the key visual information needed by the question. Samples without such evidence are discarded. The remaining samples undergo human review of question and answer correctness and joint dependence on video evidence and external knowledge. Finally, verified samples are stratified by video length and task difficulty before balanced sampling.

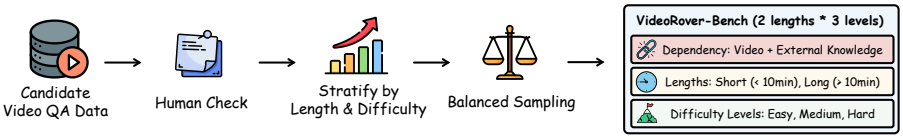
We divide videos into *short* (0–10 minutes) and *long* (>10 minutes) subsets, and further categorize questions by temporal-grounding and retrieval difficulty:

- **Easy:** the relevant visual evidence is readily localized and only a few searches are required.

[3]The DeepSeek-V4-Flash used here is the Preview version released on April 24, 2026.



(a) Automated grounding and evidence verification.



(b) Human review, stratification, and balanced sampling.

Figure 3: Construction pipeline of `VideoRover-Bench`.

- **Medium:** the evidence is more difficult to localize or the answer requires more retrieval steps.
- **Hard:** the visual evidence is elusive, or the answer requires joint reasoning over multiple video segments and search results.

Balanced sampling retains 50 examples for each of the 6 length–difficulty combinations, yielding **300 evaluation instances** in total. See Appendix A for additional statistics.

5 Model Training

We train `VideoRover` in two complementary stages. SFT establishes a reliable tool-use prior that links video observation with external retrieval, while RL improves autonomous decisions under actual tool feedback and final-answer supervision.

SFT Training. We first fine-tune the base model on the verified multi-round trajectories from Section 4.2. Unlike final-answer supervision, these trajectories demonstrate how to localize video evidence, select image or text search, inspect webpages, update the evidence state, and decide whether to continue or answer. SFT therefore provides the model with a stable behavioral initialization for the complete Video Deep Research process, including valid tool invocation and decisions based on returned evidence. Let $x_i = (V_i, Q_i)$ denote the task input and τ_i^* its target trajectory. Tool outputs are treated as environmental context and excluded from the loss over model-generated reasoning, tool calls, and answer tokens. The objective is $\mathcal{L}_{\mathrm{SFT}}(\theta) = -\frac{1}{M}\sum_{i=1}^{M} \log \pi_\theta(\tau_i^* \mid x_i)$.

RL Training. Although SFT establishes the basic research behavior, it imitates fixed paths and does not optimize success under actual tool feedback. We therefore apply Group Sequence Policy Optimization (GSPO) (Zheng et al., 2025a) to the challenging data from Section 4.3, allowing the policy to explore alternative decisions and reinforce complete observation–retrieval–verification trajectories through final-answer correctness. For each input x, the old policy samples $G = 8$ trajectories. Correct and incorrect answers receive rewards of 1 and 0, while trajectories with formatting errors, missing answers, or invalid tool calls are filtered out. We estimate trajectory-level advantages using the leave-one-out baseline $\widehat{A}_i = R_i - \frac{1}{G-1}\sum_{j\neq i} R_j$.

Let y_i be the i-th trajectory and $\rho_i(\theta)$ the sequence-level importance ratio between the current and old policies. With $p_{i,t}^{\theta} = \pi_\theta(y_{i,t} \mid x, y_{i,<t})$, we compute $\rho_i(\theta) = \exp\left(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} \log \frac{p_{i,t}^{\theta}}{p_{i,t}^{\theta_{\mathrm{old}}}}\right)$. This ratio averages log-probability changes over valid tokens within each trajectory. GSPO then applies sequence-level clipping and averages the resulting objective across trajectories:

$$\mathcal{J}_{\mathrm{GSPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G} \min\left(\rho_i(\theta)\widehat{A}_i, \mathrm{clip}(\rho_i(\theta), 1-\epsilon_{\mathrm{low}}, 1+\epsilon_{\mathrm{high}})\,\widehat{A}_i\right). \tag{4}$$

This sequence-level clipping aligns policy updates with trajectory-level rewards while reducing length bias. Together with challenging data and outcome-based rewards, GSPO turns the SFT initialization into a policy that learns when to gather evidence, revisit the video, or answer.

Table 1: Main results on VideoDR and `VideoRover-Bench`. All values are accuracy (%). S and L denote short and long videos, respectively. Avg. is the unweighted mean over VideoDR and the six `VideoRover-Bench` subsets. `VideoRover` variants are shaded in blue. Within the open-source agentic group, the best and second-best results are shown in bold and underlined, respectively. † denotes the result reported in its original paper.

Model	VideoDR	VideoRover-Bench						Avg.
	Acc.	S-Easy	S-Medium	S-Hard	L-Easy	L-Medium	L-Hard	
Proprietary Models (Direct Answer)								
GPT-5 (Singh et al., 2025)	57.00	70.00	64.00	36.00	64.00	50.00	44.00	55.00
GPT-5.2 (Singh et al., 2025)	50.00	68.00	52.00	32.00	70.00	40.00	46.00	51.14
GPT-5.4 (Singh et al., 2025)	62.00	74.00	68.00	36.00	70.00	44.00	54.00	58.29
Gemini-2.5-Flash (Comanici et al., 2025)	47.00	54.00	30.00	30.00	54.00	30.00	28.00	39.00
Gemini-2.5-Pro (Comanici et al., 2025)	54.00	66.00	54.00	30.00	62.00	34.00	50.00	50.00
Gemini-3-Flash (Pichai et al., 2025)	59.00	74.00	70.00	48.00	58.00	52.00	48.00	58.43
Gemini-3-Pro (Pichai et al., 2025)	55.00	78.00	60.00	42.00	72.00	40.00	52.00	57.00
Open-Source Models (Direct Answer)								
Qwen3-VL-8B (Bai et al., 2025a)	9.00	8.00	8.00	8.00	18.00	6.00	10.00	9.57
Qwen3-VL-30B-A3B (Bai et al., 2025a)	10.00	14.00	16.00	12.00	22.00	10.00	12.00	13.71
Qwen3.5-27B (Qwen Team, 2026a)	28.00	32.00	30.00	14.00	28.00	22.00	12.00	23.71
Qwen3.5-35B-A3B (Qwen Team, 2026a)	24.00	36.00	28.00	14.00	36.00	20.00	14.00	24.57
Qwen3.6-27B (Qwen Team, 2026b)	23.00	30.00	30.00	16.00	38.00	18.00	12.00	23.86
Qwen3.6-35B-A3B (Qwen Team, 2026c)	28.00	36.00	18.00	14.00	26.00	22.00	16.00	22.86
Open-Source Models (Agentic Tool Use)								
Qwen3-VL-8B (Bai et al., 2025a)	29.00	58.00	38.00	16.00	42.00	30.00	10.00	31.86
Qwen3-VL-30B-A3B (Bai et al., 2025a)	31.00	58.00	34.00	14.00	50.00	22.00	16.00	32.14
Qwen3.5-27B (Qwen Team, 2026a)	<u>54.00</u>	<u>78.00</u>	54.00	36.00	<u>68.00</u>	**56.00**	**42.00**	<u>55.43</u>
Qwen3.5-35B-A3B (Qwen Team, 2026a)	42.00	56.00	42.00	18.00	38.00	22.00	20.00	34.00
Qwen3.6-27B (Qwen Team, 2026b)	<u>54.00</u>	74.00	**64.00**	<u>38.00</u>	66.00	46.00	34.00	53.71
Qwen3.6-35B-A3B (Qwen Team, 2026c)	50.00	76.00	<u>62.00</u>	34.00	62.00	48.00	<u>40.00</u>	53.14
VideoSearcher-8B (Gao et al., 2026)	53.00†	64.00	46.00	28.00	54.00	36.00	22.00	43.29
`VideoRover-8B-SFT` (Ours)	39.00	58.00	44.00	30.00	54.00	32.00	18.00	39.29
`VideoRover-8B-RL` (Ours)	**56.00**	**80.00**	**64.00**	**40.00**	**70.00**	<u>52.00</u>	**42.00**	**57.71**

6 EXPERIMENT

6.1 EXPERIMENTAL SETUP

Training Details. We use Qwen3-VL-8B (Bai et al., 2025a) as the backbone of `VideoRover` and optimize it through SFT followed by RL. Due to space limitations, full implementation and training details are provided in Appendix C.1.

Evaluation Details. We evaluate Video Deep Research on VideoDR (Liu et al., 2026a) and our proposed `VideoRover-Bench`. Beyond this setting, we assess general video understanding on MLVU (Test Set) (Zhou et al., 2025), Video-MMMU (Hu et al., 2026), Video-MME (Fu et al., 2025), and LVBench (Wang et al., 2025). On the Video Deep Research benchmarks, we compare proprietary models under direct answering with open-source models under both direct answering and ReAct-style agentic tool use (Yao et al., 2023). All agentic models can autonomously invoke the same tools under a shared evaluation protocol. On the general video benchmarks, we compare against a range of representative open-source video-reasoning methods and backbone models. Additional implementation and evaluation settings are provided in Appendix C.2.

6.2 MAIN RESULTS

Video Deep Research. Table 1 compares direct answering with agentic tool use. Proprietary models achieve strong performance even without external tools. Their larger capacity and broad parametric world knowledge allow them to answer many questions directly from the observed video and internal knowledge. In contrast, open-source models perform considerably worse under direct answering, indicating that limited parametric knowledge and passive video observation are insufficient for questions that jointly require precise temporal evidence and external information.

Equipping the open-source models with the same tools as `VideoRover` consistently yields substantial improvements over their direct-answer counterparts. This result demonstrates the value of turning video question answering into an active research process that can revisit relevant segments, verify visual entities, retrieve missing knowledge, and inspect source webpages. However,

Table 2: Results on general video understanding benchmarks. The best and second-best results are shown in bold and underlined, respectively.

Model	MLVU ≈ 651s	Video-MMMU ≈ 506s				Video-MME (w/o subtitles) ≈ 1018s				LVBench ≈ 4101s
	M-Avg	Adapt.	Compr.	Percep.	Avg.	Short	Medium	Long	Overall	Avg.
Video-R1 (Feng et al., 2026)	52.3	46.5	<u>49.0</u>	63.6	53.0	71.3	62.8	53.2	62.4	39.8
VideoChat-R1 (Li et al., 2025)	<u>56.5</u>	45.5	43.6	62.0	50.4	72.5	63.1	53.8	63.1	41.1
VideoRFT (Wang et al., 2026)	51.5	45.5	45.0	62.0	50.8	70.8	63.5	54.2	62.8	39.6
VideoChat-R1.5 (Yan et al., 2026)	55.9	<u>47.3</u>	45.0	64.0	52.1	71.1	<u>64.6</u>	52.5	62.7	40.6
LongVT (Yang et al., 2026)	48.6	45.1	45.0	52.6	47.6	66.6	57.6	50.7	58.3	40.2
Video-MTR (Xie et al., 2026)	50.9	40.5	39.6	58.3	46.1	71.6	61.1	50.1	60.9	39.8
VideoTemp-o3 (Liu et al., 2026b)	50.5	41.9	41.3	58.3	47.2	70.1	61.0	49.8	60.3	41.5
Video-o3 (Zeng et al., 2026)	55.1	<u>47.3</u>	46.3	63.3	52.3	**74.1**	63.8	50.4	62.8	41.5
VideoZoomer (Ding et al., 2026)	51.9	39.7	43.3	60.6	47.9	73.0	61.8	56.3	63.7	42.9
Qwen2.5-VL-7B (Bai et al., 2025b)	50.9	44.8	47.6	61.0	51.1	68.3	59.2	51.3	59.6	37.2
Qwen3-VL-8B (Bai et al., 2025a)	55.5	<u>47.3</u>	45.6	<u>69.0</u>	<u>54.0</u>	72.6	62.3	<u>57.2</u>	<u>64.0</u>	<u>43.7</u>
VideoRover-8B-RL (Ours)	**57.3**	**62.0**	**63.3**	**70.6**	**65.3**	<u>73.8</u>	**67.8**	**59.1**	**66.9**	**45.0**

the performance differences among agentic models also show that tool access alone is insufficient. Effective Video Deep Research further requires the model to select and compose tools according to the evolving evidence state rather than invoking them as an unstructured pipeline.

VideoRover-8B-RL achieves the strongest overall performance among the evaluated open-source agentic models and leads or matches them on most subsets. Despite its 8B backbone, it reaches a level comparable to the strongest proprietary direct-answer models and surpasses newer, larger open-source models operating with the same tools, as well as the similarly sized VideoSearcher-8B baseline (Gao et al., 2026). This result shows that our trajectory construction and task-specific training can compensate for limited model scale by transforming missing parametric knowledge into explicit evidence acquisition. The consistent improvement from VideoRover-8B-SFT to VideoRover-8B-RL further confirms that RL strengthens autonomous long-horizon coordination beyond imitation of the synthesized trajectories.

General Video Benchmarks. Table 2 shows that VideoRover-8B-RL achieves the best overall performance on MLVU, Video-MME, and LVBench. Its improvements remain consistent across benchmarks with different video durations and reasoning requirements. These results demonstrate that VideoRover not only excels at Video Deep Research but also generalizes effectively to broader video understanding tasks.

Unlike MLVU, Video-MME, and LVBench, which mainly assess reasoning over instance-specific evidence contained in each video, Video-MMMU is more knowledge-intensive: it treats professional educational videos as knowledge sources and evaluates whether a model can perceive, comprehend, and adapt the acquired knowledge (Hu et al., 2026; Fu et al., 2026). Answering its questions therefore requires not only locating relevant instructional content but also understanding the introduced concepts and applying them to new problems. By coupling repeated video inspection with image and text search and webpage browsing, VideoRover can retrieve external knowledge that complements the video and use it to interpret the observed content. It consequently obtains particularly large improvements in comprehension and adaptation while also achieving the best overall result. This demonstrates that the research capabilities learned for Video Deep Research can enhance knowledge acquisition from videos, beyond improving instance-specific video understanding alone.

6.3 Ablation Studies

Table 3 shows that VideoRover performs best across video lengths and difficulty levels, confirming that each tool contributes to the research process. Removing crop_video causes a degradation across the benchmark, demonstrating the importance of revisiting candidate segments at higher temporal density and visual resolution before retrieval. Removing image_search also reduces performance, with an evident effect on long-video subsets. This supports its role in confirming visual entities from localized keyframes and preventing uncertain recognition from propagating into subsequent searches. Among the individual tools, removing text_search produces the largest overall decline, indicating that many questions require the model to connect video-grounded cues with external background or factual knowledge. The removal of visit further weakens performance, showing that search-result summaries alone are often insufficient and that opening source web-

Table 3: Ablation results on VideoDR and VideoRover-Bench. All values are accuracy (%). Each row removes one tool from VideoRover-8B-RL, except web, which jointly removes all external retrieval tools. S and L indicate short and long videos, respectively. Bold and underlined values denote the best and second-best results.

Model	VideoDR	VideoRover-Bench						Avg.
	Acc.	S-Easy	S-Medium	S-Hard	L-Easy	L-Medium	L-Hard	
VideoRover-8B-RL	**56.00**	**80.00**	**64.00**	**40.00**	**70.00**	**52.00**	**42.00**	**57.71**
− crop_video	<u>45.00</u>	74.00	52.00	<u>34.00</u>	64.00	42.00	36.00	<u>49.57</u>
− image_search	43.00	<u>76.00</u>	<u>54.00</u>	<u>34.00</u>	56.00	44.00	36.00	49.00
− text_search	24.00	40.00	26.00	20.00	30.00	16.00	10.00	23.71
− visit	41.00	<u>76.00</u>	44.00	30.00	<u>68.00</u>	<u>46.00</u>	<u>38.00</u>	49.00
− web	13.00	16.00	18.00	14.00	16.00	10.00	10.00	13.86

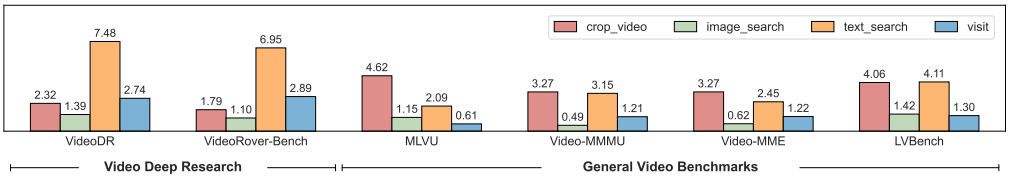


Figure 4: Average per-example tool calls by VideoRover-8B-RL across the two benchmark types.

pages is necessary to extract detailed evidence. The web ablation jointly removes image_search, text_search, and visit, causing a substantially larger degradation than removing any single tool and highlighting their complementarity. Together, these results show how VideoRover coordinates its tools based on the available evidence: crop_video grounds the research process in relevant visual evidence, image_search verifies visual entities, text_search retrieves missing knowledge, and visit consolidates detailed evidence for the final answer.

6.4 Tool-Use Analysis

Figure 4 shows distinct tool-use patterns. On VideoDR and VideoRover-Bench, VideoRover invokes text_search, image_search, and visit more frequently, reflecting a greater need for external information. In contrast, on the general video benchmarks, it makes fewer external-retrieval calls but uses crop_video more often, indicating that these tasks depend more on locating and inspecting evidence within the input video. The gap is most pronounced for text_search, highlighting the central role of external knowledge in Video Deep Research. This contrast shows that VideoRover does not follow a fixed tool-use pipeline, but adapts its actions to each question's demand for external knowledge, balancing in-video exploration with open-world retrieval.

7 Conclusion

We presented VideoRover, a unified framework that iteratively coordinates video observation and open-world retrieval, conditioning each action on the observations accumulated so far. To train and evaluate it, we synthesized verified SFT and RL trajectories spanning temporal localization, multimodal search, webpage browsing, and evidence synthesis, and built VideoRover-Bench, stratified by video duration and research difficulty. On VideoDR and VideoRover-Bench, VideoRover-8B-RL performs on par with proprietary models in the direct-answer setting and outperforms larger open-source models given the same tools. Results on general video benchmarks demonstrate broader generalization, while ablation and tool-use analyses validate complementary tool design and adaptive coordination between video inspection and external retrieval.

Limitations. Despite its strong Video Deep Research performance, VideoRover incurs additional inference overhead because multi-round tool interactions increase both the number of reasoning steps and the overall response time compared with direct answering. Moreover, due to the substantial cost of SFT and RL, we evaluate only Qwen3-VL-8B as the backbone, leaving its scaling behavior and transferability to other model families unexplored. Future work could develop more efficient tool-use and stopping strategies and extend training to backbones of different sizes and architectures. Finally, the model depends on the availability and quality of external search results, motivating further study of robustness to noisy, incomplete, or unavailable web evidence.

AI USE STATEMENT

We used generative AI tools to polish the manuscript, assist with code development, and synthesize training and benchmark data. The roles of the models used in data construction and trajectory synthesis are described in the main text. We did not treat AI-generated content as authoritative or allow generative AI to make final decisions about experimental findings or scientific conclusions. All AI-assisted outputs were rigorously reviewed by the authors: manuscript revisions were checked for accuracy and consistency, generated code was inspected and tested, and synthetic data were filtered and verified using the procedures described in the paper, including human review where applicable. No AI-generated output was included without author approval. The authors take full responsibility for the final text, code, data, claims, and conclusions presented in this work.

ETHICS STATEMENT

This work does not involve human-subject experiments or the collection of private or sensitive personal data. All webpages and YouTube videos used for data construction were publicly accessible at the time of collection and were used solely for research purposes. Source references are retained where applicable to support traceability. Public web content may nevertheless contain factual errors or societal biases. Our multi-stage filtering and human verification reduce but cannot completely eliminate these risks.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we include the complete training and evaluation code in the supplementary material. Section 4 documents the full data acquisition and processing pipeline, including source collection, visual grounding, question and trajectory synthesis, filtering, and benchmark construction. Section 5 specifies the SFT and RL objectives, while Section 6 and Appendix C provide the experimental protocol, training hyperparameters, inference budgets, baseline configurations, and evaluation settings. Appendix G further provides the complete prompt templates used throughout data construction, trajectory synthesis, and evaluation. Together, these resources provide the code, data-processing procedures, and configurations needed to reproduce the reported experiments.

REFERENCES

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16776–16809, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Shuang Chen, Kaituo Feng, Hangting Chen, Wenxuan Huang, Dasen Dai, Quanxin Shou, Yunlong Lin, Xiangyu Yue, Shenghua Gao, and Tianyu Pang. Opensearch-vl: An open recipe for frontier multimodal search agents. *arXiv preprint arXiv:2605.05185*, 2026.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5558–5570, 2022.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for

long videos. In *International Conference on Learning Representations*, volume 2025, pp. 18227–18246, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Yang Ding, Xin Lai, Yizhen Zhang, Wei Li, Ruihang Chu, and Yujiu Yang. Videozoomer: Reinforcement-learned temporal focusing for long video reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 20087–20111, 2026.

Zhen Fang, Yu Zeng, Wenxuan Huang, Yiming Zhao, Shiting Huang, Tianfei Ren, Qi Lu, Qingnan Ren, Qisheng Su, Lionel Z Wang, et al. Video-deepresearch: Towards the next-generation multimodal deepresearch agent. *arXiv preprint arXiv:2608.03979*, 2026.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Lin Fu, Zheyuan Yang, Yang Wang, Tingyu Song, Arman Cohan, and Yilun Zhao. Videokr: Towards knowledge- and reasoning-intensive video understanding. In *Forty-third International Conference on Machine Learning*, 2026.

Zhenkun Gao, Yicheng Bao, Jinlong Peng, Xueheng Li, Theo Huang, Bangwei Liu, Kunquan Li, Zhenye Gan, Tao Hu, Chengjun Xie, et al. Videosearcher: Empowering video deep research with multi-tool agentic reasoning via reinforcement learning. *arXiv preprint arXiv:2607.02927*, 2026.

Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Kuan Li, Yida Zhao, et al. Webwatcher: Breaking new frontiers of vision-language deep research agent. In *International Conference on Learning Representations*, volume 2026, pp. 61240–61271, 2026.

Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. Framethinker: Learning to think with long videos via multi-turn frame spotlighting. In *International Conference on Learning Representations*, volume 2026, pp. 74904–74933, 2026.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, and Guohai Xu. Deepeyesv2: Toward agentic multimodal model. In *International Conference on Learning Representations*, volume 2026, pp. 114851–114872, 2026.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Xiang Yue, Bo Li, Yuanhan Zhang, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multidisciplinary professional videos. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27798–27828, 2026.

Wenxuan Huang, Yu Zeng, Qiuchen Wang, Zhen Fang, Shaosheng Cao, Zheng Chu, Qingyu Yin, Shuang Chen, Zhenfei Yin, Lin Chen, et al. Vision-deepresearch: Incentivizing deepresearch capability in multimodal large language models. 2026.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Guankai Li, Jiabin Chen, Yi Xu, Xichen Zhang, and Yuan Lu. Hypereyes: Dual-grained efficiency-aware reinforcement learning for parallel multimodal search agents. *arXiv preprint arXiv:2605.07177*, 2026a.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. Webthinker: Empowering large reasoning models with deep research capability. *Advances in Neural Information Processing Systems*, 38:120091–120131, 2026b.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.

Zhengyang Liang, Yan Shu, Xiangrui Liu, Minghao Qin, Kaixin Liang, Nicu Sebe, Zheng Liu, and Lizi Liao. Video-browser: Towards agentic open-web video browsing. *arXiv preprint arXiv:2512.23044*, 2025.

Chengwen Liu, Xiaomin Yu, Zhuoyue Chang, Zhe Huang, Shuo Zhang, Heng Lian, Jisheng Dang, Rui Xu, Sen Hu, Jianheng Hou, et al. Watching, reasoning, and searching: A video deep research benchmark on open web for agentic video reasoning. *arXiv preprint arXiv:2601.06943*, 2026a.

Wenqi Liu, Yunxiao Wang, Shijie Ma, Meng Liu, Qile Su, Tianke Zhang, Haonan Fan, Changyi Liu, Kaiyu Jiang, Jiankang Chen, et al. Videotemp-o3: Harmonizing temporal grounding and video understanding in agentic thinking-with-videos. In *Forty-third International Conference on Machine Learning*, 2026b.

Jiahao Meng, Yue Tan, Qi Xu, Kuan Gao, Weisong Liu, Yanwei Li, Jason Li, Lingdong Kong, Haochen Wang, Qianyu Zhou, et al. Watch, remember, reason: Human-view video understanding with mllms. *arXiv preprint arXiv:2606.07433*, 2026.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Sundar Pichai, Demis Hassabis, and Koray Kavukcuoglu. A new era of intelligence with Gemini 3, November 2025. URL `https://blog.google/products-and-platforms/products/gemini/gemini-3`.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL `https://qwen.ai/blog?id=qwen3.5`.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026b. URL `https://qwen.ai/blog?id=qwen3.6-27b`.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026c. URL `https://qwen.ai/blog?id=qwen3.6-35b-a3b`.

Yuxiang Shen, Hailong Huang, Zhenkun Gao, Xueheng Li, Man Zhou, Chengjun Xie, Haoxuan Che, Xuanhua He, and Jie Zhang. Lookwise: Knowing when and where to look for fine-grained visual reasoning in multimodal large language models. *arXiv preprint arXiv:2603.00171*, 2026.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Zhengliang Shi, Yiqun Chen, Haitao Li, Weiwei Sun, Shiyu Ni, Yougang Lyu, Run-Ze Fan, Bowen Jin, Yixuan Weng, Minjun Zhu, et al. Deep research: A systematic survey. *arXiv preprint arXiv:2512.02038*, 2025.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. Openthinkimg: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *Advances in neural information processing systems*, 38:4350–4376, 2026.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22958–22967. IEEE, 2025.

Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, et al. Searching for best practices in retrieval-augmented generation. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 17716–17736, 2024.

Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. Mmsearch-r1: Incentivizing lmms to search. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2456–2487, 2026.

Yuan Xie, Tianshui Chen, Zheng Ge, and Lionel Ni. Video-MTR: Reinforced multi-turn reasoning for long video understanding. In *Forty-third International Conference on Machine Learning*, 2026.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Renjun Xu and Jingwen Peng. A comprehensive survey of deep research: Systems, methodologies, and applications. *arXiv preprint arXiv:2506.12594*, 2025.

Ziang Yan, Xinhao Li, Yinan He, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *Advances in Neural Information Processing Systems*, 38:119152–119184, 2026.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, et al. Longvt: Incentivizing" thinking with long videos" via native tool calling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33816–33826, 2026.

Huanjin Yao, Qixiang Yin, Min Yang, Ziwang Zhao, Yibo Wang, Haotian Luo, Jingyi Zhang, and Jiaxing Huang. Mm-deepresearch: A simple and effective multimodal agentic search baseline. *arXiv preprint arXiv:2603.01050*, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Xiangyu Zeng, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, Zikang Wang, Changlian Ma, Qingyu Zhang, Zizheng Huang, Kun Ouyang, Tianxiang Jiang, et al. Video-o3: Native interleaved clue seeking for long video multi-hop reasoning. In *Forty-third International Conference on Machine Learning*, 2026.

Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32903–32914, 2026a.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.

Peng Zhang, Guanghao Zhang, Wanggui He, Longxiang Zhang, Mushui Liu, Yan Xia, Zhenhao Peng, Weilong Dai, Jinlong Liu, Haobing Tang, et al. Dynframe: Adaptive reasoning-driven multimodal framework with dynamic frame augmentation for complex video understanding. *arXiv preprint arXiv:2605.26680*, 2026b.

YiFan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, Haonan fan, Kaibing Chen, Jiankang Chen, Haojie Ding, Kaiyu Tang, Zhang Zhang, Liang Wang, Fan Yang, Tingting Gao, and Guorui Zhou. Thyme: Think beyond images. In *International Conference on Learning Representations*, volume 2026, pp. 79536–79572, 2026c.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, et al. Swift: a scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 29733–29735, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025a.

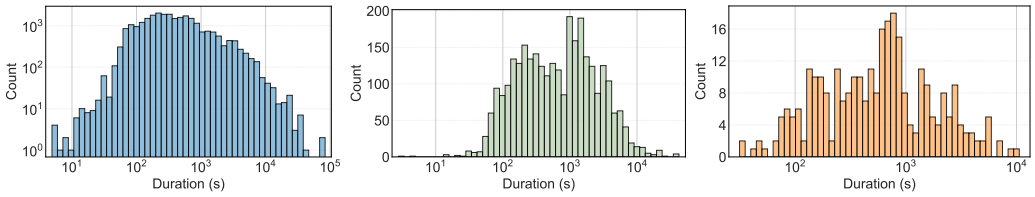
Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 414–431, 2025b.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, and Chao Shen. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 126775–126798, 2026.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701. IEEE, 2025.

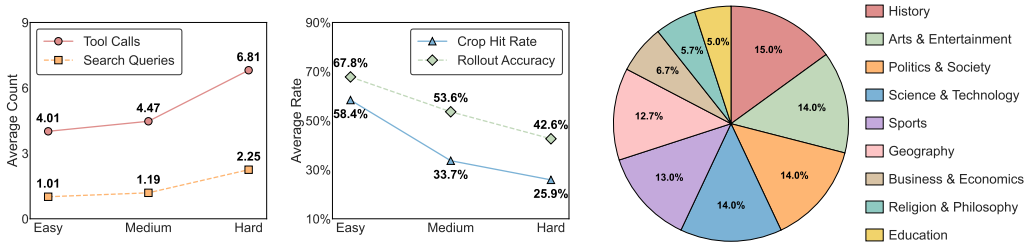
A Data and Benchmark Statistics

To better characterize our training data and benchmark, we provide additional statistics on video duration, difficulty stratification, and knowledge-domain coverage in this section.



(a) SFT data. (b) RL data. (c) `VideoRover-Bench`.

Figure 5: Video-duration distributions of the SFT data, RL data, and `VideoRover-Bench`.



(a) Interaction counts. (b) Validation rates. (c) Knowledge domains.

Figure 6: Statistics of `VideoRover-Bench`: (a) average tool calls and search queries by difficulty, (b) crop hit rate and rollout accuracy by difficulty, and (c) knowledge-domain distribution.

Figure 5 shows that the SFT data, RL data, and `VideoRover-Bench` span a broad range of video durations. This diversity supports training and evaluation across temporal scales, allowing the model to learn from and be assessed on both short and long videos. Figure 6a and Figure 6b report statistics computed from rollouts generated by the SFT model. The crop hit rate measures whether `crop_video` captures the visual evidence required to answer the question. Average tool calls and search queries increase from easy to hard, while both the crop hit rate and rollout accuracy decrease, demonstrating a clear and well-grounded difficulty stratification in the candidate benchmark data. Subsequent human review further helps ensure the correctness and validity of the retained examples. Figure 6c further shows that `VideoRover-Bench` maintains broad and relatively balanced coverage across nine knowledge domains, supporting diverse evaluation beyond a single topic.

B Tool Implementations

We implement the four tools behind the shared interface used during training and evaluation. The implementations of `text_search`, `image_search`, and `visit` follow Vision-DeepResearch (Huang et al., 2026).

Video cropping. The `crop_video` tool validates the requested temporal boundaries and uses Decord[4] by default to decode the corresponding video segment and return uniformly sampled frames for dense visual inspection.

Text search. The `text_search` tool is implemented with the Serper API[5]. It submits one or more queries to Google Search and returns the titles, URLs, dates, and snippets of up to 10 organic results per query. These entries serve as candidate sources for subsequent verification with `visit`.

[4]https://github.com/dmlc/decord

[5]https://serper.dev/

Image search. The `image_search` tool extracts the selected video frame, uploads it to object storage, and uses the same Serper API for reverse-image retrieval. It then visits up to three matched pages to return evidence relevant to the current goal.

Webpage visit. The `visit` tool accepts up to five URLs and retrieves their contents through Jina Reader[6]. It then uses Qwen3-VL-30B-A3B Bai et al. (2025a) as an LLM-based extractor to produce concise, goal-conditioned evidence and summaries. Search and webpage results are cached, and failed requests are retried to improve execution stability.

C Experimental Details

C.1 Training Configuration

We use *ms-swift* (Zhao et al., 2025) for SFT and *VeRL* (Sheng et al., 2025) for RL. The learning rates are set to 2×10^{-5} and 1×10^{-6}, with batch sizes of 256 and 16, for SFT and RL, respectively. During RL, *vLLM* (Kwon et al., 2023) samples 8 rollout trajectories per training example, and each rollout can make at most 25 tool calls. We use GPT-4o-mini (Hurst et al., 2024) as the outcome judge, assigning binary rewards based on final-answer correctness. For the initial video observation, we sample at 1 FPS, retain at most 256 frames, and limit each frame to a maximum resolution of 224×224 pixels. The `crop_video` tool instead samples at 2 FPS and returns at most 32 frames, each with a maximum resolution of 448×448 pixels. All experiments are conducted on 4 servers, each equipped with 8 NVIDIA H800 GPUs and 2 TB of memory.

C.2 Evaluation Configuration

Video Deep Research. The proprietary direct-answer baselines comprise GPT-5, GPT-5.2, GPT-5.4 (Singh et al., 2025), Gemini-2.5-Flash, Gemini-2.5-Pro (Comanici et al., 2025), Gemini-3-Flash, and Gemini-3-Pro (Pichai et al., 2025). The open-source baselines comprise Qwen3-VL-8B, Qwen3-VL-30B-A3B (Bai et al., 2025a), Qwen3.5-27B, Qwen3.5-35B-A3B (Qwen Team, 2026a), Qwen3.6-27B (Qwen Team, 2026b), and Qwen3.6-35B-A3B (Qwen Team, 2026c), each evaluated under both direct answering and ReAct-style agentic tool use (Yao et al., 2023). We additionally include VideoSearcher-8B (Gao et al., 2026), a specialized Video Deep Research agent trained for multi-tool reasoning. In the agentic setting, the ReAct baselines, `VideoRover`-8B-SFT, and `VideoRover`-8B-RL can make at most 50 tool calls per example. We cap the initial video input at 64 frames for consistent comparison, with all other settings following the training configuration.

General Video Understanding. We compare with specialized video-reasoning and thinking-with-video methods, including Video-R1 (Feng et al., 2026), VideoChat-R1 (Li et al., 2025), VideoChat-R1.5 (Yan et al., 2026), VideoRFT (Wang et al., 2026), LongVT (Yang et al., 2026), Video-MTR (Xie et al., 2026), VideoTemp-o3 (Liu et al., 2026b), Video-o3 (Zeng et al., 2026), and Video-Zoomer (Ding et al., 2026). We additionally report Qwen2.5-VL-7B (Bai et al., 2025b) and Qwen3-VL-8B (Bai et al., 2025a) as backbone baselines. For a consistent comparison, all general-video baselines use the same video preprocessing as `VideoRover`, sampling at 1 FPS and retaining at most 256 frames for the initial input. All remaining settings follow their official configurations.

D Fine-Grained Video Deep Research Results

To complement the aggregate results in Table 1, we report domain- and difficulty-level performance on VideoDR, together with domain-level performance on `VideoRover-Bench`. These breakdowns assess whether the overall gains of `VideoRover` generalize across question categories and difficulty levels.

As shown in Table 4, RL matches or improves upon SFT across every VideoDR domain and improves performance at every difficulty level. `VideoRover`-8B-RL achieves the strongest overall result among open-source models and performs particularly well on the high-difficulty subset, indicating that long-horizon optimization benefits questions requiring more involved grounding and

[6] https://jina.ai/

Table 4: Domain- and difficulty-stratified results on VideoDR. All values are accuracy (%). Within the open-source agentic group, the best and second-best results in each column are shown in bold and underlined, respectively.

Model	Domain						Difficulty			Overall
	Daily Life	Econ.	Culture	Tech.	History	Geo.	Low	Mid	High	
Proprietary Models (Direct Answer)										
GPT-5 (Singh et al., 2025)	51.52	56.25	73.33	46.67	63.64	60.00	65.62	58.33	46.88	57.00
GPT-5.2 (Singh et al., 2025)	48.48	50.00	60.00	33.33	54.55	60.00	59.38	47.22	43.75	50.00
GPT-5.4 (Singh et al., 2025)	60.61	56.25	73.33	66.67	63.64	50.00	71.88	66.67	46.88	62.00
Gemini-2.5-Flash (Comanici et al., 2025)	45.45	43.75	66.67	40.00	45.45	40.00	46.88	58.33	34.38	47.00
Gemini-2.5-Pro (Comanici et al., 2025)	54.55	50.00	66.67	33.33	63.64	60.00	59.38	55.56	46.88	54.00
Gemini-3-Flash (Pichai et al., 2025)	42.42	68.75	66.67	60.00	90.91	50.00	75.00	61.11	40.62	59.00
Gemini-3-Pro (Pichai et al., 2025)	51.52	68.75	60.00	46.67	54.55	50.00	71.88	55.56	37.50	55.00
Open-Source Models (Direct Answer)										
Qwen3-VL-8B (Bai et al., 2025a)	12.12	6.25	13.33	6.67	9.09	0.00	15.62	5.56	6.25	9.00
Qwen3-VL-30B-A3B (Bai et al., 2025a)	12.12	6.25	13.33	6.67	18.18	0.00	15.62	5.56	9.38	10.00
Qwen3.5-27B (Qwen Team, 2026a)	33.33	25.00	40.00	20.00	27.27	10.00	31.25	30.56	21.88	28.00
Qwen3.5-35B-A3B (Qwen Team, 2026a)	21.21	37.50	26.67	20.00	18.18	20.00	40.62	22.22	9.38	24.00
Qwen3.6-27B (Qwen Team, 2026b)	21.21	37.50	13.33	13.33	18.18	40.00	28.12	19.44	21.88	23.00
Qwen3.6-35B-A3B (Qwen Team, 2026c)	33.33	37.50	33.33	20.00	18.18	10.00	34.38	27.78	21.88	28.00
Open-Source Models (Agentic Tool Use)										
Qwen3-VL-8B (Bai et al., 2025a)	27.27	43.75	26.67	33.33	18.18	20.00	46.88	25.00	15.62	29.00
Qwen3-VL-30B-A3B (Bai et al., 2025a)	21.21	25.00	33.33	33.33	45.45	50.00	46.88	27.78	18.75	31.00
Qwen3.5-27B (Qwen Team, 2026a)	**54.55**	50.00	**66.67**	**53.33**	54.55	40.00	**84.38**	47.22	31.25	54.00
Qwen3.5-35B-A3B (Qwen Team, 2026a)	39.39	37.50	40.00	46.67	54.55	40.00	56.25	33.33	37.50	42.00
Qwen3.6-27B (Qwen Team, 2026b)	**54.55**	50.00	53.33	**53.33**	54.55	**60.00**	**84.38**	44.44	34.38	54.00
Qwen3.6-35B-A3B (Qwen Team, 2026c)	45.45	**56.25**	60.00	**53.33**	54.55	30.00	65.62	**52.78**	31.25	50.00
VideoSearcher-8B (Gao et al., 2026)	–	–	–	–	–	–	–	–	–	53.00
VideoRover-8B-SFT (Ours)	42.42	37.50	53.33	33.33	36.36	20.00	59.38	33.33	25.00	39.00
VideoRover-8B-RL (Ours)	51.52	**56.25**	53.33	46.67	**90.91**	50.00	78.12	41.67	**50.00**	**56.00**

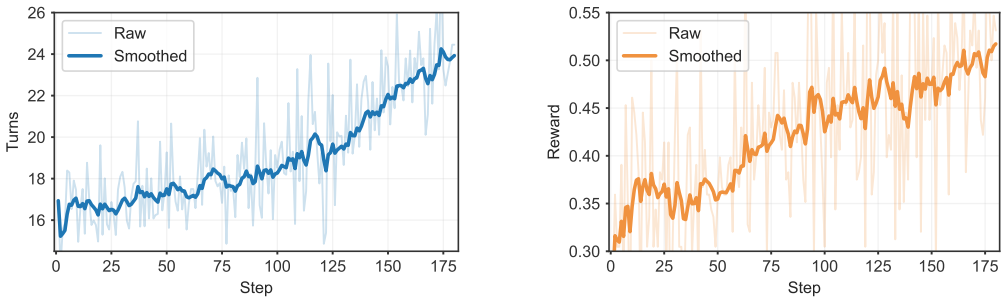
Table 5: Domain-stratified results on `VideoRover-Bench`. All values are accuracy (%). Within the open-source agentic group, the best and second-best results in each column are shown in bold and underlined, respectively.

Model	A&E	B&E	Edu.	Geo.	Hist.	P&S	R&P	S&T	Sports	Overall
Proprietary Models (Direct Answer)										
GPT-5 (Singh et al., 2025)	57.14	55.00	60.00	44.74	66.67	47.62	70.59	61.90	38.46	54.67
GPT-5.2 (Singh et al., 2025)	52.38	50.00	60.00	50.00	62.22	40.48	70.59	42.86	48.72	51.33
GPT-5.4 (Singh et al., 2025)	59.52	45.00	66.67	47.37	62.22	59.52	64.71	57.14	58.97	57.67
Gemini-2.5-Flash (Comanici et al., 2025)	40.48	25.00	40.00	31.58	48.89	40.48	52.94	35.71	25.64	37.67
Gemini-2.5-Pro (Comanici et al., 2025)	50.00	50.00	53.33	42.11	66.67	35.71	64.71	52.38	38.46	49.33
Gemini-3-Flash (Pichai et al., 2025)	59.52	50.00	46.67	52.63	71.11	50.00	70.59	64.29	53.85	58.33
Gemini-3-Pro (Pichai et al., 2025)	61.90	70.00	60.00	60.53	66.67	45.24	64.71	47.62	51.28	57.33
Open-Source Models (Direct Answer)										
Qwen3-VL-8B (Bai et al., 2025a)	9.52	0.00	13.33	10.53	11.11	7.14	5.88	14.29	10.26	9.67
Qwen3-VL-30B-A3B (Bai et al., 2025a)	9.52	20.00	13.33	13.16	22.22	11.90	11.76	9.52	17.95	14.33
Qwen3.5-27B (Qwen Team, 2026a)	23.81	25.00	20.00	18.42	26.67	21.43	35.29	23.81	17.95	23.00
Qwen3.5-35B-A3B (Qwen Team, 2026a)	23.81	20.00	20.00	26.32	33.33	23.81	23.53	33.33	10.26	24.67
Qwen3.6-27B (Qwen Team, 2026b)	28.57	20.00	20.00	28.95	31.11	19.05	29.41	23.81	12.82	24.00
Qwen3.6-35B-A3B (Qwen Team, 2026c)	21.43	25.00	26.67	21.05	22.22	14.29	35.29	28.57	15.38	22.00
Open-Source Models (Agentic Tool Use)										
Qwen3-VL-8B (Bai et al., 2025a)	26.19	35.00	66.67	28.95	20.00	30.95	58.82	33.33	30.77	32.33
Qwen3-VL-30B-A3B (Bai et al., 2025a)	16.67	55.00	53.33	31.58	28.89	33.33	52.94	23.81	33.33	32.33
Qwen3.5-27B (Qwen Team, 2026a)	**61.90**	40.00	**80.00**	42.11	60.00	59.52	**70.59**	45.24	**56.41**	55.67
Qwen3.5-35B-A3B (Qwen Team, 2026a)	26.19	60.00	46.67	23.68	31.11	28.57	47.06	30.95	30.77	32.67
Qwen3.6-27B (Qwen Team, 2026b)	59.52	50.00	66.67	50.00	**64.44**	50.00	58.82	35.71	**56.41**	53.67
Qwen3.6-35B-A3B (Qwen Team, 2026c)	54.76	55.00	73.33	42.11	55.56	57.14	41.18	**57.14**	51.28	53.67
VideoSearcher-8B (Gao et al., 2026)	35.71	30.00	66.67	39.47	44.44	45.24	52.94	35.71	41.03	41.67
VideoRover-8B-SFT (Ours)	35.71	40.00	66.67	31.58	37.78	33.33	52.94	40.48	41.03	39.33
VideoRover-8B-RL (Ours)	50.00	**65.00**	**80.00**	**57.89**	62.22	**61.90**	64.71	47.62	53.85	**58.00**

retrieval. Its gains across the six domains show that the improvement is not confined to a single type of knowledge. The large gain in the history domain may partly reflect our question-construction process, which frequently asks about background and historical information. We report only the overall VideoSearcher-8B result from its paper, as its fine-grained predictions are unavailable.

Table 5 presents the domain-level breakdown on `VideoRover-Bench`. `VideoRover`-8B-RL consistently outperforms its SFT counterpart across all nine domains and achieves the best overall open-source performance, including the best or tied-best results in several domains. This broad improvement supports the effectiveness of RL in coordinating video inspection and external retrieval across diverse knowledge demands. Aggregating the same predictions over the six duration–difficulty subsets yields the results reported in Table 1.

E RL Training Dynamics



(a) Interaction turns per rollout.

(b) Training reward.

Figure 7: RL training dynamics of `VideoRover`. Light curves show raw measurements, and dark curves show smoothed trends.

As shown in Figure 7, both the smoothed reward and interaction length increase after the SFT cold start. Because the reward depends on final-answer correctness rather than trajectory length, their joint growth suggests that longer rollouts support sustained evidence acquisition rather than merely adding redundant tool calls. This trend matches the design of our RL data and indicates that RL strengthens the long-horizon coordination of video localization, multimodal search, and webpage browsing. Together with the ablation results, it shows that `VideoRover` benefits from learning to compose tools across research trajectories, rather than simply having access to them.

F Qualitative Cases

We present five qualitative cases in Figures 8–12 to illustrate how `VideoRover` coordinates temporal grounding and open-world retrieval across diverse questions. These cases show the model localizing and revisiting relevant video segments, selecting image or text search according to the observed evidence, and browsing source webpages to resolve ambiguity and verify the final answer.

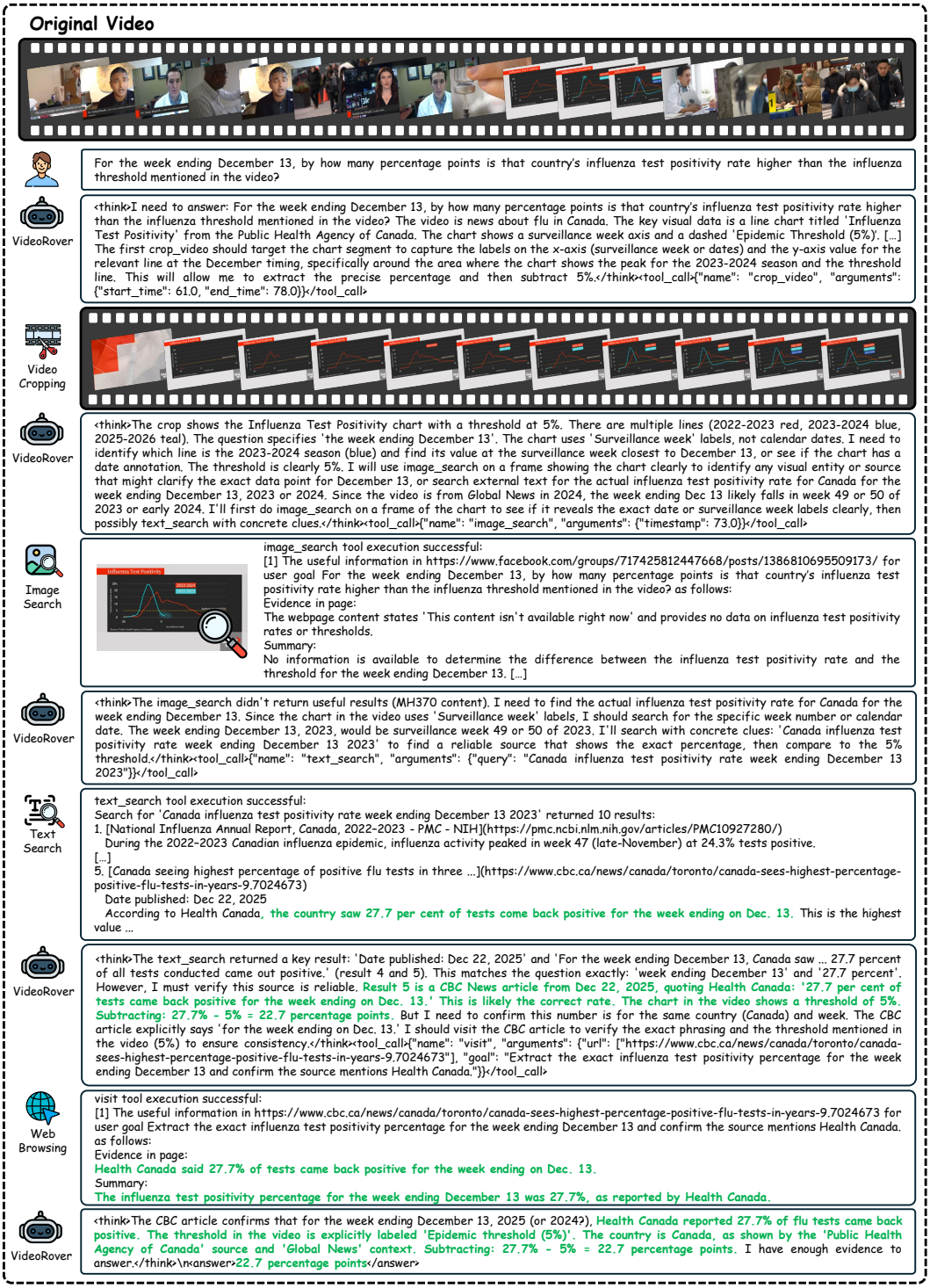


Figure 8: Qualitative case on comparing Canada's influenza test positivity rate with the threshold shown in the video. `VideoRover` localizes the relevant chart, supplements an uninformative image search with text search, and verifies the reported rate through webpage browsing to derive a difference of 22.7 percentage points.

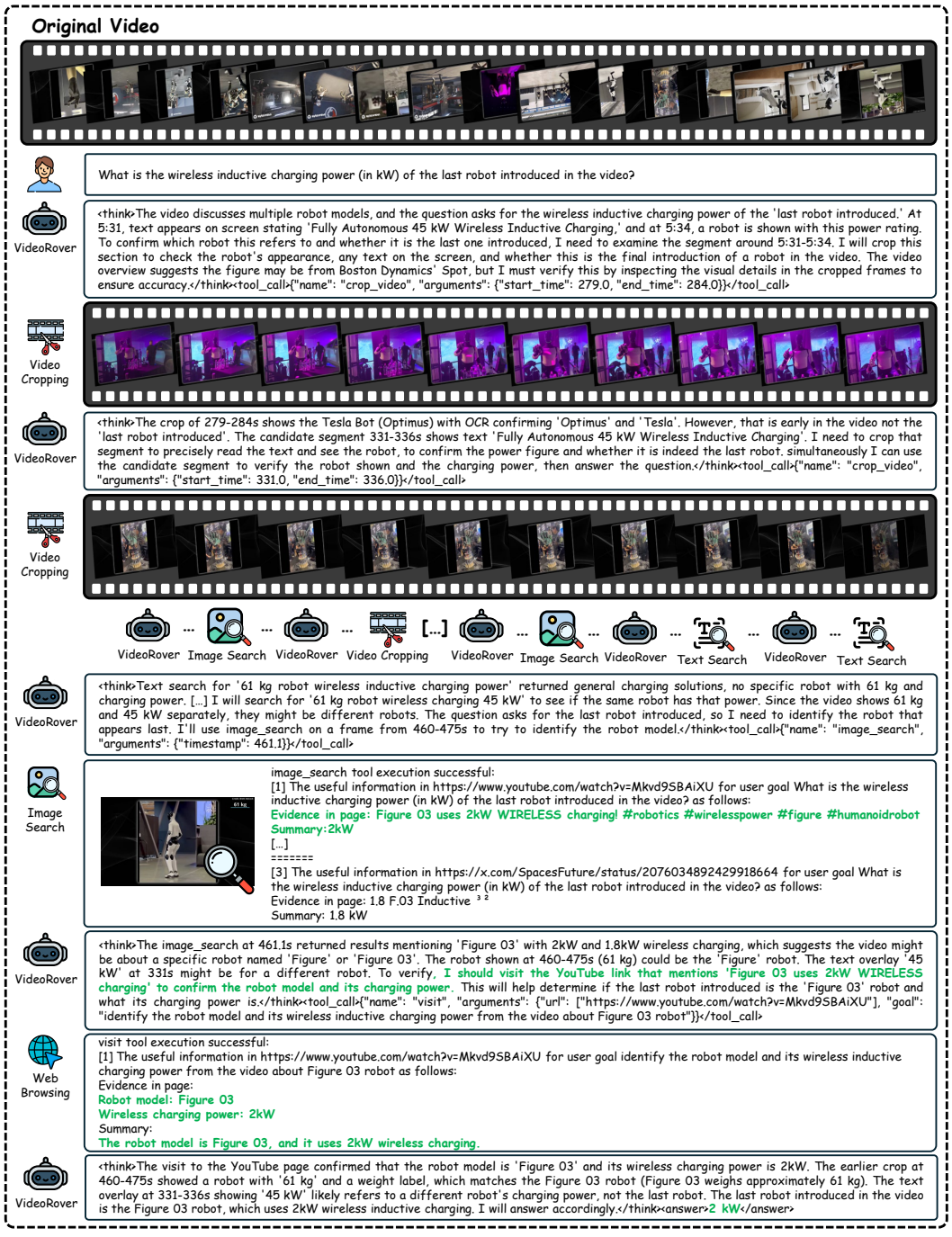


Figure 9: Qualitative case on identifying the wireless charging power of the last robot introduced in a video. `VideoRover` revisits multiple candidate segments, identifies Figure 03 through image search, and verifies its 2 kW specification on the retrieved webpage.

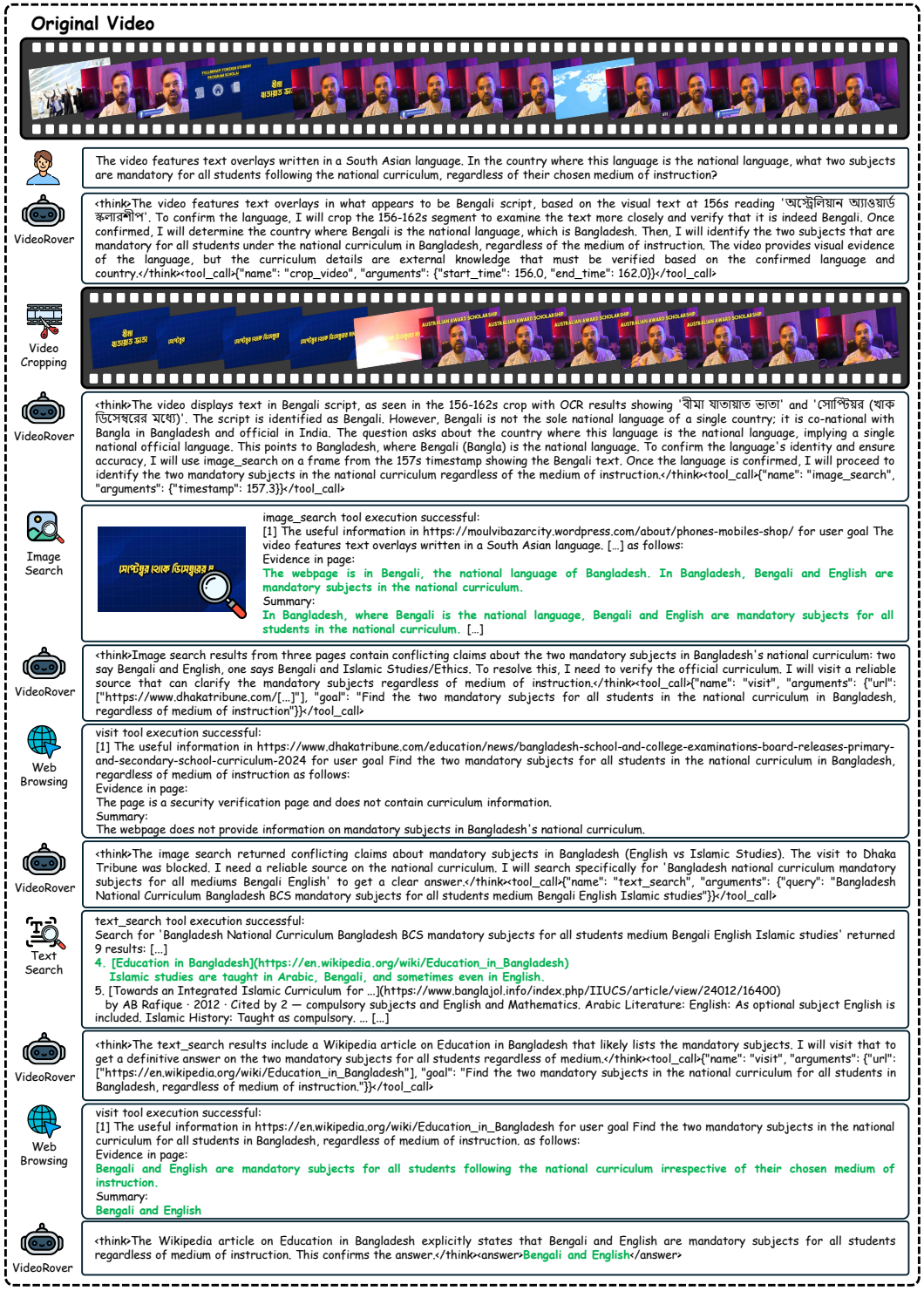


Figure 10: Qualitative case on identifying mandatory subjects from Bengali text shown in a video. `VideoRover` grounds the script through video cropping and image search, resolves conflicting retrieval results with text search and webpage browsing, and concludes that Bengali and English are mandatory.

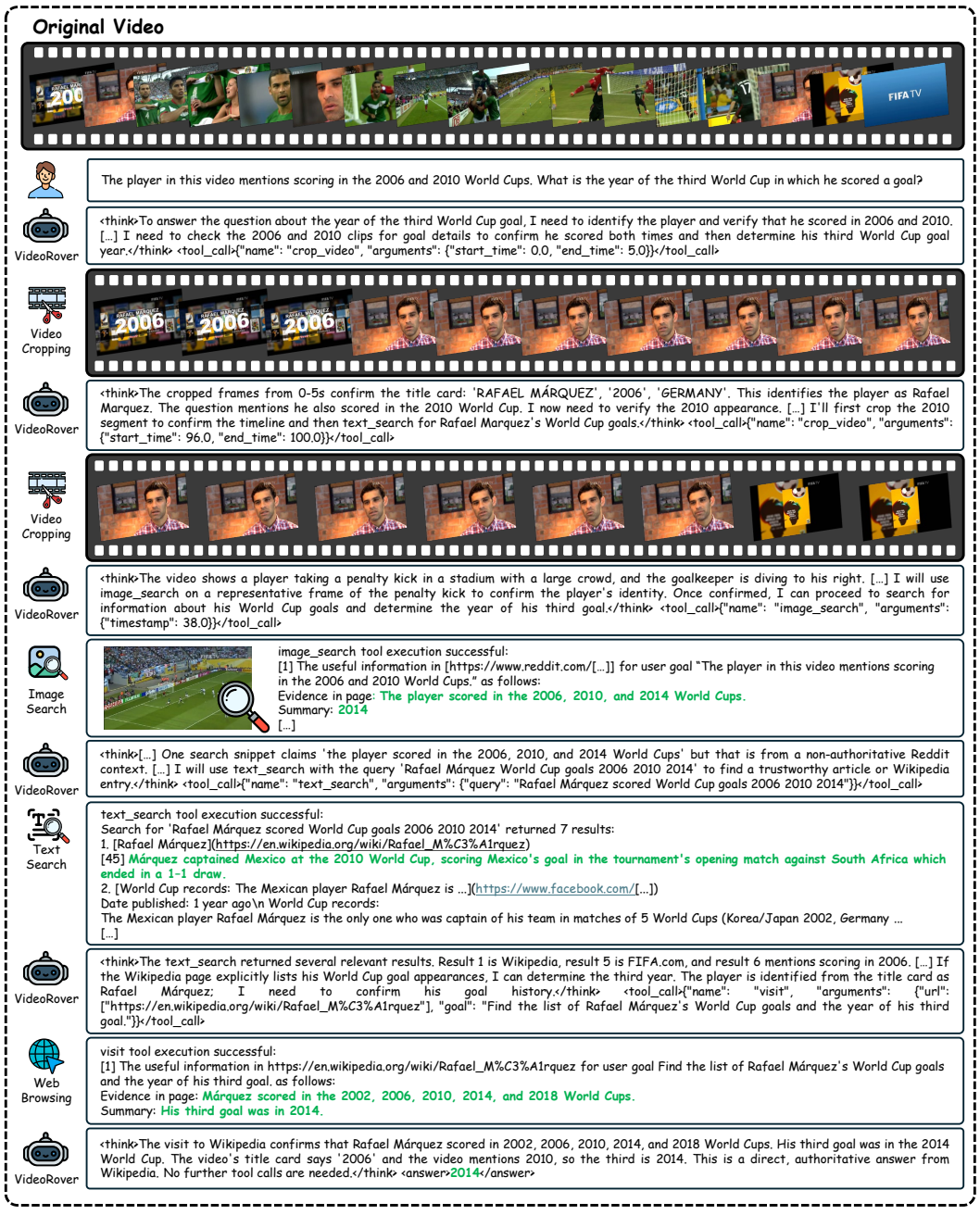


Figure 11: Qualitative case on identifying the year of Rafael Márquez's third World Cup goal. `VideoRover` crops the interview and match segments, combines image search with text search, and visits a retrieved webpage to verify the final answer.

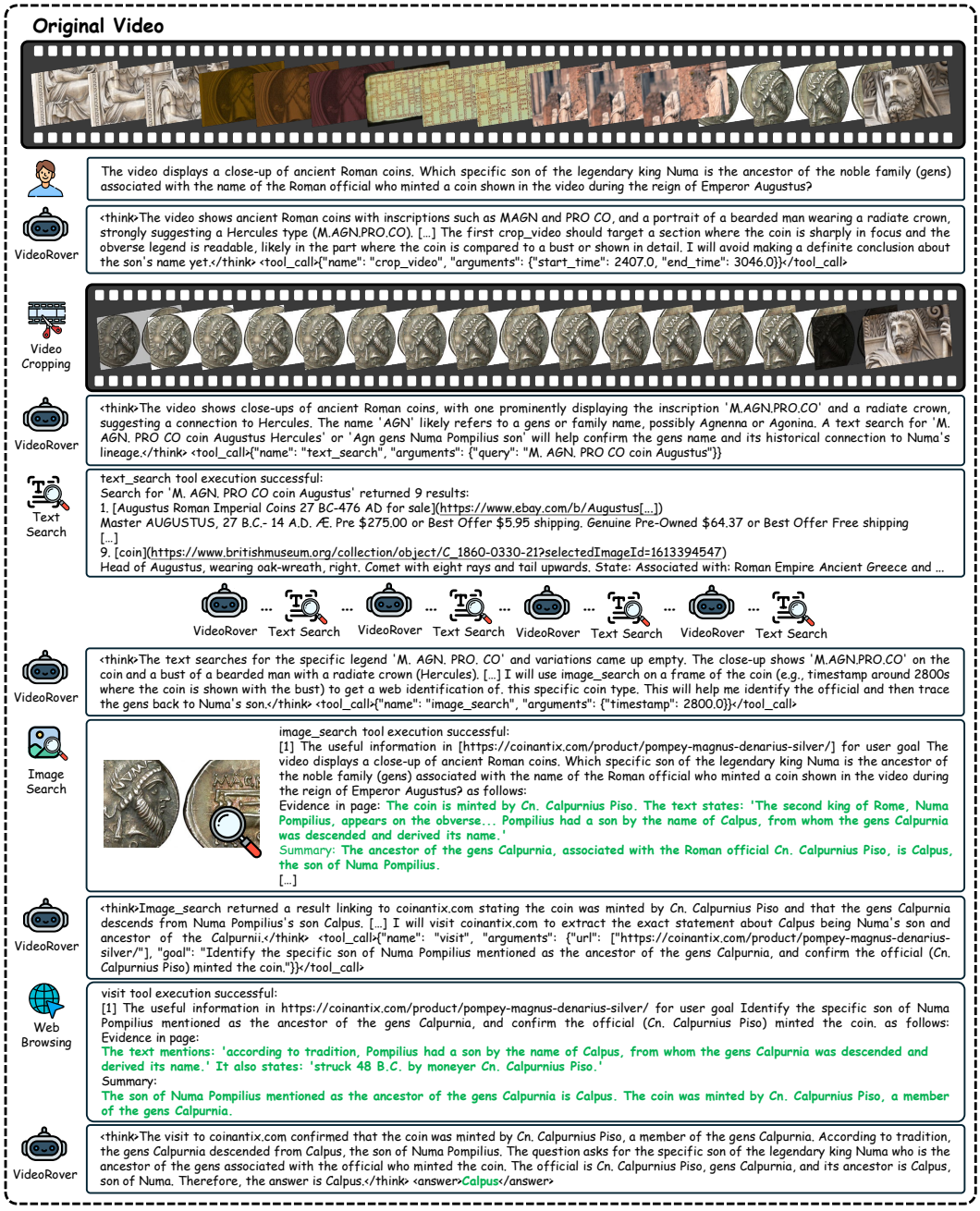


Figure 12: Qualitative case on tracing the ancestry of the gens associated with an ancient Roman coin. After localizing the coin in the video, `VideoRover` revises unsuccessful text-search attempts with image search and webpage browsing, ultimately identifying Calpus as the relevant son of Numa Pompilius.

G PROMPT TEMPLATES

This section documents the prompts used to construct questions, synthesize tool-use trajectories, and validate supervision in Section 4, ordered by their actual execution. The displayed system and user messages preserve their runtime separation. Braced text denotes an instance-dependent placeholder populated with the current question, video metadata, accumulated evidence, or interaction history.

G.1 MODEL SYSTEM PROMPT

Figure 13 defines the assistant role, four tools from Section 3, interaction protocol, and output format shared by training and evaluation. Every turn contains evidence-specific reasoning followed by exactly one tool call or final answer. The prompt also requires an initial `crop_video`, and requires a page visit before external facts become final evidence, aligning synthesized supervision with inference-time behavior.

G.2 QUESTION CONSTRUCTION

After knowledge collection and video–target grounding, Figures 14 and 15 receive the grounded video and supporting material and jointly produce a question, answer, timestamps, reasoning trace, and verifiable source cues. The template enforces joint dependence on video and external knowledge, and favors facts not recoverable from the question alone. Figure 16 then returns keep, rewrite, or discard. It removes leaked identities, timestamps, and references to supplementary documents while preserving the visual anchor.

G.3 TRAJECTORY GENERATION

Following Section 4.2, DeepSeek-V4-Flash serves as the research planner and Qwen3.5-27B as the video observer. The implementation separately tracks visual evidence, tool evidence, tool history, and the assistant-facing trajectory. The planner chooses the next evidence gap, whereas the observer selects exact timestamps. The templates follow three stages:

1. **Roles and initialization.** Sparse frames yield a timeline, clues, uncertain entities, and candidate crops (Figure 17). The planner proposes a visual sub-goal (Figure 20); the observer selects its interval (Figure 19) and parses dense frames for visual and OCR evidence (Figure 18).
2. **Evidence-guided tool use.** The planner iteratively crops, identifies an entity, searches, visits, or answers from the accumulated state (Figure 21). Runtime checks reject repeated or poorly grounded actions and require visited-page evidence beyond search snippets; Figure 22 extracts concise evidence from each page.
3. **Answer synthesis and filtering.** Once sufficient evidence is available, the final-answer template synthesizes a concise answer from video observations and successfully visited sources, explicitly resolving conflicting evidence when necessary (Figure 23). If the interaction budget is reached first, a final answer is generated only when the required visual and webpage evidence has been verified. The trajectory-quality judge checks answer support, tool validity, redundancy, and visual grounding (Figure 25), while the answer-equivalence judge tolerates harmless wording differences but remains strict about factual values, dates, and counts (Figure 26). For retained trajectories, a final cleanup pass rewrites only the surface form of the reasoning, removing hidden orchestration language without adding facts or changing the underlying evidence chain (Figure 24).

Trajectory System Prompt

You are a multimodal deep research assistant for video question answering.

Your job is to construct a rigorous video deepsearch reasoning trajectory. You must analyze the video, ground visual entities carefully, use tools to verify external facts, and answer the user's rewritten question accurately.

Core Requirements.

- At the beginning, form an overall plan from the question and visible video content, then call `crop_video` to locate the key segment. The first assistant turn must contain the plan inside `<think>`, followed by a `crop_video` tool call.
- Every `<think>` must contain concrete reasoning for this exact sample: visible video clues, uncertainty, why the next tool is needed, what sub-goal the tool should verify, or how gathered evidence changes the plan.
- If a visual identity is uncertain, use `image_search` on a relevant frame before making a firm identity claim.
- `image_search` is reverse-image/web lookup, not OCR, zoom, caption extraction, or date reading. For on-screen text, production credits, dates, captions, tables, or small text, use `crop_video` and frame parsing.
- `text_search` queries must be fine-grained sub-goals using concrete video/OCR/image clues, not copies of the whole question.
- When `text_search` returns relevant results, call `visit` on the most relevant source before reasoning to a final external-fact answer.
- If `visit` fails, is blocked, says the page cannot verify the fact, or contradicts the current hypothesis, do not answer from search snippets. Search again or visit a better source.
- Treat tool outputs as evidence. Do not invent facts, URLs, search results, webpage content, or tool outputs.

Available Tools.

1. **`crop_video`**

   `<tool_call>{"name":"crop_video","arguments":{"start_time":0.0, "end_time":10.0}}</tool_call>`
2. **`image_search`**

   `<tool_call>{"name":"image_search","arguments":{"timestamp":12.3}}</tool_call>`
3. **`text_search`**

   `<tool_call>{"name":"text_search","arguments":{"query":"focused query with concrete clues"}}</tool_call>`
4. **`visit`**: Use 1 to 5 explicit URLs selected from search results according to the actual verification need.

   `<tool_call>{"name":"visit","arguments":{"url":["https://example.com"], "goal":"specific fact to verify"}}</tool_call>`

Output Format. Each assistant turn must output exactly one of the following:

`<think>...</think><tool_call>...</tool_call>`
or
`<think>...</think><answer>...</answer>`

Never output a tool call and final answer in the same turn.

Figure 13: Assistant-facing system prompt defining the interaction protocol, available tools, and required output format.

Question Generating

System Prompt

You are an expert dataset designer specializing in complex, multi-modal Video-QA (Video Question Answering).

User Prompt

Your task is to design high-quality, challenging questions based on the provided Video and Supplementary Materials.

Design Criteria.

1. **Visual Anchoring (MOST IMPORTANT):** The question MUST require observing specific visual elements in the video to answer—such as identifying an object, reading on-screen text, recognizing a person/place/logo, observing an action, or noticing a spatial arrangement. If the question can be answered by simply knowing the topic (e.g., "What year did X happen?") without needing to watch the video at all, it is INVALID. Ask yourself: "Would someone need to watch the video to even understand what this question is asking about?" If no, redesign the question.
2. **Question Phrasing—Video-Only Perspective (STRICTLY ENFORCED):** The question must be phrased as if asked by someone who ONLY watched the video and has NO knowledge of any supplementary materials. The question must NEVER contain ANY of the following: "supplementary materials," "the materials," "the article," "the document," "the text," "Wikipedia," "according to," "as described in," "as mentioned in," "the provided information," or any similar reference to external text sources. The question should ONLY reference things that are visually or audibly observable in the video. Use phrases like "the building shown at 0:45," "the logo visible on the uniform," or "the device being demonstrated"—always anchoring to something visually observable.
3. **Cross-Modal Dependency:** The question CANNOT be answered by watching the video alone OR by reading the materials alone. The solver must: (a) first identify a specific visual element from the video, then (b) look up detailed information about that element in the supplementary materials. Both steps are essential.
4. **Knowledge Dependency (Anti-Memorization):** The answer must rely on highly specific information found ONLY in the provided materials. It should be obscure enough that an LLM cannot answer it from parametric memory.
5. **Complex Reasoning:** The question must require multi-hop reasoning. The solver should: extract visual cues from the video $\rightarrow$ identify what to search for $\rightarrow$ find relevant details in external knowledge $\rightarrow$ synthesize the final answer.

Common Mistakes to Avoid.

- **BAD:** "What language family does the word mentioned in the video belong to?" This is a pure text/knowledge question; no visual observation is needed.
- **BAD:** "What is the population of the city discussed in the video?" Knowing the city name is enough; no video watching is required.
- **BAD:** "According to the supplementary materials, what year was the bridge built?" This directly references the materials. NEVER do this.
- **BAD:** "Based on the provided text, how many species are endangered?" This references "provided text." NEVER do this.
- **BAD:** "The video mentions that X comes from Y. What is Z?" This rephrases material content as if the video said it. NEVER do this.
- **GOOD:** "What is the architectural style of the building with the red dome visible at the start of the video?" This requires identifying the specific building visually, then looking up its style.
- **GOOD:** "The instrument the performer switches to at 2:30—in what year was this model first manufactured?" This requires visual identification of the instrument model, then external lookup.

Figure 14: Prompt used to generate video-grounded, knowledge-dependent questions and reference answers.

Question Generating (Continued)

Output Format. You must output the result strictly in JSON format with the following keys:

- `Question`: The designed question. It must require visual observation of the video to understand and answer. NO reference to supplementary materials.
- `Answer`: The concise final answer.
- `Timestamps`: The key video segments required to answer, formatted as `[[start_seconds, end_seconds], ...]`. Times must be integers or floats in **seconds** (e.g., `[[123, 142]]`), NOT in "mm:ss" or "hh:mm:ss" format.
- `Reasoning_process`: A step-by-step explanation showing how visual cues from the video lead to a lookup in the materials, and then to the answer.
- `Cues`: The exact key text snippets extracted from the supplementary materials that are essential for answering the question.

Examples of Expected Output.

```
{
  "Question": "What is the museum registration number of the artifact displayed
closest to the main entrance shown in the video?",
  "Answer": "WB. 67",
  "Timestamps": [[23, 56], [78, 83]],
  "Reasoning_process": "First, identify the specific museum via visual cues in the
video (e.g., building exterior, signs). Then, search external sources to find the
museum's floor plan, locate the exhibit closest to the entrance among the 'don't miss'
recommendations, and extract its accession number WB.67.",
  "Cues": "According to the museum's floor plan, the 'Don't miss' exhibit situated
immediately to the right of the main entrance is the Holy Thorn Reliquary, bearing the
registration number WB. 67."
}
```

```
{
  "Question": "What is the rated generating capacity per unit in MW for the turbines
shown being installed in the video?",
  "Answer": "15",
  "Timestamps": [[45, 57]],
  "Reasoning_process": "First, localize the video clip where the specific power units
appear and identify the project name or unit model from the visual context. Next,
search for the technical specifications of those turbines to determine the capacity is
15 MW.",
  "Cues": "The Phase II expansion project utilizes the newly acquired Siemens SGT-400
gas turbines. Each of these units has a rated generating capacity of 15 MW under
standard conditions."
}
```

```
{
  "Question": "In the video, there is a close-up of a yellow medical card with a ship
name heavily crossed out. In what year did that ocean liner officially enter
service?",
  "Answer": "1890",
  "Timestamps": [[35, 79], [104, 122]],
  "Reasoning_process": "First, find the yellow medical/inspection card in the video
and carefully analyze the crossed-out text to identify the name of the ocean liner (SS
Majestic). Then, search maritime historical records to find the exact year this
specific ship entered service.",
  "Cues": "The SS Majestic, frequently noted in quarantine records with its name
struck through due to changing port regulations, was launched in 1889 and officially
entered passenger service in the spring of 1890."
}
```

Critical Rule. After your thinking, output ONLY a single valid JSON object. Do NOT include markdown formatting around the JSON. Timestamps must be in seconds, for example:

```
{"Question":"...", "Answer":"...", "Timestamps":[[123,142]],
"Reasoning_process":"...", "Cues":"..."}
```

Now, I will provide you the video and supplementary material:
Video: [The video is provided above]
Supplementary Material: {`material`}

Figure 15: Continuation of the question-generation prompt.

Question Filter

System Prompt

You are an expert at rewriting Video-QA questions.

User Prompt

Your task is to evaluate and rewrite the given question according to strict rules.

Original Question: {question}
Original Answer: {answer}

Rules for Rewriting.

1. **Remove specific video details that give away the answer.** The question should NOT tell the solver exactly what is shown in the video (e.g., specific timestamps like "at 04:24," specific names/labels read from the video like "identified as 'Gori Fortress'," or detailed descriptions of video content). Instead, refer to visual elements vaguely so the solver must watch the video to find them. For example:
   - **BAD:** "The fortress identified as 'Gori Fortress' at 04:24 overlooks the town of Gori. Which Catholic religious order ..."
   - **GOOD:** "Which Catholic religious order was conducting a major mission in the town overlooked by the fortress shown in the video in 1633?"
2. **Remove all references to supplementary materials / references / documents.** If the question mentions "supplementary materials," "the provided reference list," "the article," "the document," "Wikipedia," "according to the text," "the provided information," or any similar phrase, rewrite it to remove these references. If the question fundamentally depends on referencing a specific document/list/table that cannot be replaced with a video-based question, mark it as DISCARD.
3. **Ensure the question requires watching the video.** If the question can be fully answered just by reading the text of the question itself (without watching any video), it must be DISCARDED. The rewritten question should require the solver to identify something visual in the video first.
4. **Preserve the core intent.** The rewritten question should still ask for the same answer. Do not change what is being asked—only change HOW it is phrased.

When to Discard.

- The question explicitly references a "reference list," "citation," "provided table," or "supplementary materials" and cannot be rewritten without them.
- The question is purely text/knowledge-based with no meaningful visual anchor in a video (e.g., "What language family does the word X belong to?").
- The question reveals so much video content in text form that it becomes a pure text question.

Output Format. After your thinking, output ONLY one of the following:

- Rewrite: {"rewritten_question":"the new question text", "action":"rewrite"}
- Keep: {"rewritten_question":"the original question text unchanged", "action":"keep"}
- Discard: {"rewritten_question":"", "action":"discard", "reason":"brief reason"}

Examples.
Original: "The fortress identified as 'Gori Fortress' at 04:24 overlooks the town of Gori. Which Catholic religious order was conducting a major mission in this town in 1633?"
Output: {"rewritten_question":"Which Catholic religious order was conducting a major mission in the town overlooked by the fortress shown in the video in 1633?", "action":"rewrite"}
Original: "The video explicitly names the philosophical work written by Marcus Aurelius. In the provided reference list, locate the specific citation for this work where the translator is identified as 'Farquharson' and the section number contains the digit 34. What are the two page numbers associated with this entry?"
Output: {"rewritten_question":"", "action":"discard", "reason":"Question depends on a specific reference list that cannot be replaced with video-based observation"}
Original: "The video demonstrates the operator priming the pan before winding the mainspring. Does this order of operations align with the standard procedure for the wheellock mechanism?"
Output: {"rewritten_question":"Does the order of operations shown in the video align with the standard procedure for the wheellock mechanism?", "action":"rewrite"}
Original: "The monk in the video explains that Japanese shrines worship invisible deities, whereas temples worship Buddha statues. What are the main equivalents to these deities on Okinawa referred to as in English?"
Output: {"rewritten_question":"", "action":"discard", "reason":"Question can be answered from text alone without visual observation"}
Original: "The video displays the date 256 BC as the end of the Zhou Dynasty at 01:43. Identify the ruler from the historical list whose reign ended in this same year."
Output: {"rewritten_question":"", "action":"discard", "reason":"Question references a historical list from supplementary materials and gives away the key information in text"}

Figure 16: Prompt used to rewrite or discard questions according to visual-grounding and knowledge-dependency requirements.

Roles and Initialization: Video Overview

System Prompt

You are a careful video observation model. Return JSON only. Do not solve external facts.

User Prompt

Describe the video for a downstream text reasoning model.
Question:
{question}
Video duration: {duration}s.
Input timestamp hints, if any:
{timestamp_hints}
Return valid JSON only:

```
{
  "summary": "overall visual description; separate visible facts from inferred
identities",
  "timeline": [
    {"time_range": "0-5s", "visual": "what is visible", "ocr": "readable text if
any"}
  ],
  "key_clues": ["objects, landmarks, signs, flags, names, dates, brands, maps, people,
actions"],
  "uncertain_entities": ["visual entities that need crop or image_search confirmation;
include entities whose identity is inferred rather than explicitly readable"],
  "candidate_crop_segments": [
    {"start_time": 12.0, "end_time": 22.0, "reason": "why this segment matters"}
  ]
}
```

Figure 17: Prompt used to summarize sparsely sampled frames and propose candidate crop segments.

Roles and Initialization: Cropped-Frame Analysis

System Prompt

You are a careful frame/OCR parser. Return JSON only. Do not invent external facts.

User Prompt

Analyze these frames extracted from the video for a downstream text reasoner.
Question:
{question}
Source tool: {tool_name}
Frame index to timestamp mapping:
{frame_timestamps}
Return valid JSON only:

```
{
  "summary": "specific observations useful for answering the question; distinguish
direct OCR/visible content from inferred identity",
  "frames": [
    {"index": 0, "timestamp": 0.0, "visual": "visible content", "ocr": "readable
text",
      "entities": ["possible names"], "uncertainty": "what remains uncertain"}
  ],
  "key_visual_entities": ["entities or visual candidates grounded by these frames"],
  "search_clues": ["concrete clues that should be used in image_search or
text_search"],
  "recommended_next_checks": ["image_search/text_search/visit sub-goals if needed"]
}
```

Figure 18: Prompt used to parse densely sampled crop or image frames for visual and OCR evidence.

Roles and Initialization: Temporal Localization

System Prompt

You select video timestamps for tool calls. Return JSON only.

User Prompt

You are the video analysis model responsible for selecting timestamps for visual tools.
Question:
{question}
Video duration: {duration}s.
The text reasoner has decided the next tool should be {tool_name}. The text reasoner may have included timestamp arguments, but those arguments are only a weak hint. You must choose the actual timestamp or time span by analyzing the video and the evidence below.
Reasoner think:
{reasoner_think}
Reasoner proposed arguments, if any:
{reasoner_arguments}
Input timestamp hints, if any:
{timestamp_hints}
Candidate crop segments from your earlier video observation:
{candidate_crop_segments}
Already inspected crop spans:
{used_crop_spans}
Initial video observation:
{video_overview}
Visual evidence from cropped/image frames:
{visual_evidence}
Tool evidence:
{tool_evidence}
Current trajectory:
{trajectory}
Tool history: {tool_history}
Rules.

- Select timestamps from the video itself. Do not let the text reasoner determine timestamps.
- Prefer input timestamp hints when they align with the visible content and the current unresolved sub-goal.
- Do not default to the opening 0–10 seconds unless the question or video evidence genuinely points there.
- For crop_video, select a compact segment that can expose the needed visual/OCR evidence; avoid spans already inspected unless no better option exists.
- For image_search, select a single clear representative frame of the uncertain visual entity. Do not choose a frame merely because it contains text/OCR.
- Do not solve the final external fact. Only choose the timestamp/time span.

For crop_video, return valid JSON only:

```
{"start_time": {default_start}, "end_time": {default_end},
 "reason": "why this segment best targets the unresolved visual evidence"}
```

For image_search, return valid JSON only:

```
{"timestamp": {default_timestamp},
 "reason": "why this frame is the best representative image for reverse-image
identification"}
```

Figure 19: Prompt used by the vision model to select crop spans and image-search timestamps.

Roles and Initialization: First Crop Planning

System Prompt

You write high-quality SFT assistant reasoning. Return JSON only.

User Prompt

You are constructing the first assistant turn of a VideoRAG SFT trajectory.
Write as the multimodal assistant itself. Do not mention DeepSeek, a judge model, a vision parser, or hidden helper models.
Question:
{question}
Video overview from visual observation:
{video_overview}
Video duration: {duration}s.
Task.

- Make a concrete overall plan grounded in the visible video clues.
- The first tool must be crop_video.
- Explain what visual/OCR evidence the first crop_video should target, but do not decide exact timestamps in the reasoning text.
- Do not state the final answer in this first think.
- Do not present an inferred visual identity as certain before the crop; phrase it as a hypothesis if needed.

Return valid JSON only:

```
{"think": "specific plan and what the first crop should inspect"}
```

Action Schema Used by the Step Planner.
Return valid JSON only in one of these forms:
For a tool:

```
{
  "think": "concrete reasoning for the next tool",
  "action": {"name": "image_search|text_search|visit|crop_video", "arguments":
{...}},
  "answer": ""
}
```

For final answer:

```
{
  "think": "evidence synthesis",
  "action": null,
  "answer": "concise direct answer"
}
```

Tool Argument Requirements.

- crop_video: arguments may be {} or {"target": "what visual/OCR evidence to inspect"}; exact start_time/end_time may be omitted.
- image_search: arguments may be {} or {"target": "uncertain visual entity to identify"}; exact timestamp may be omitted.
- text_search: {"query": "focused query using concrete visual/OCR/search clues"}.
- visit: {"url": ["https://..."], "goal": "specific fact to verify"}; url must contain 1 to 5 explicit URLs selected as needed.

Figure 20: Prompt and output schema used to construct the first trajectory turn.

Evidence-Guided Tool Use: Next-Step Planning

System Prompt

You are a rigorous VideoRAG trajectory planner. Return JSON only.

User Prompt

You are generating the next assistant turn for a VideoRAG SFT trajectory.
Write as the multimodal assistant itself. Do not mention DeepSeek, judge model, VLM parser, helper model, or that another model produced observations.
Question: {question}
Video duration: {duration}s
Initial video overview: {video_overview}
Visual evidence from cropped/image frames: {visual_evidence}
Tool evidence: {tool_evidence}
Current SFT trajectory context: {trajectory}
Tool history: {tool_history}
Policy note: {policy_note}
Decision Rules.

- Prefer image_search before committing to an uncertain landmark/object/person/product/robot/building/artwork/ship/logo identity.
- If the policy note asks for image_search, comply before text_search.
- For crop_video and image_search, decide whether the tool is needed and what visual target it should inspect; do not make timestamp selection the substance of the reasoning.
- Use image_search only for reverse-image/web identification of a visual entity. Never use it to read OCR, zoom in, extract production credits, captions, dates, years, or small text.
- For visible text/date/credits/OCR questions, use crop_video around the likely timestamp and rely on visual frame parsing. If text remains unreadable, crop a narrower/relevant segment, not image_search.
- Use text_search for focused sub-goals, not the whole question.
- text_search must include concrete visual anchors from crop/frame evidence when those anchors exist, such as visible names, flags, logos, OCR terms, places, or product names.
- Search snippets are recall evidence, not final evidence. If a text_search returned relevant URLs and no successful visit has verified them yet, choose visit before final answer.
- If visit is blocked, says it cannot verify the fact, returns no useful content, or contradicts the current hypothesis, do not answer from snippets. Choose another visit URL or a refined text_search.
- Visit goals must be specific and should ask the page to identify/extract a fact. Do not phrase the goal as "verify that X is Y" when X/Y are still hypotheses.
- When visit evidence contradicts visual evidence or earlier assumptions, explicitly resolve the conflict and prefer the stronger direct evidence. Do not ignore visit output.
- If evidence is still insufficient, choose a useful next tool. If enough evidence is gathered, give the final answer.
- Keep the think concrete and evidence-based. No generic workflow filler.
- Do not invent tool outputs or facts.
- Maximum remaining tool calls: {remaining_tool_calls}.

Return JSON using the action schema in Figure 20.

Figure 21: Prompt used to select each subsequent tool call or produce an evidence-grounded answer.

Evidence-Guided Tool Use: Web Evidence Extraction

System Prompt

You are a helpful assistant.

User Prompt

Please process the following webpage content and user goal to extract relevant information:
Webpage Content.
{webpage_content}
User Goal.
{goal}
Return a valid JSON object only with keys rational, evidence, and summary. Keep evidence and summary concise and directly relevant.

Invalid-JSON Repair Suffix.
Your previous output was not valid JSON:
{previous_output}
Return valid JSON only. Do not use markdown or code fences.

Figure 22: Prompt used by the visit tool to extract webpage evidence and the suffix used to repair malformed JSON responses.

Answer Synthesis: Final Answer

System Prompt

You synthesize evidence into a final SFT answer. Return JSON only.

User Prompt

Force a final answer for this VideoRAG SFT trajectory.
Write as the multimodal assistant itself. Do not mention helper models.
Question: {question}
Initial video overview: {video_overview}
Visual evidence: {visual_evidence}
Tool evidence: {tool_evidence}
Trajectory: {trajectory}
Strict Final-Answer Rules.

- Do not base the answer on text_search snippets alone when a visit failed or contradicted the hypothesis.
- If a successful visit gives a specific value, do not ignore it in favor of an earlier guess or snippet.
- If direct crop/OCR evidence and page evidence conflict, explain which source answers the actual question wording and why.
- The final answer must be supported by the listed evidence; otherwise say the evidence is insufficient instead of inventing.

Return valid JSON only:

```
{
  "think": "concise evidence synthesis from video observations and successful
visit/page evidence; mention unresolved conflicts if any",
  "answer": "direct final answer only"
}
```

Figure 23: Prompt used to synthesize a supported final answer from the collected evidence.

Answer Synthesis: Reasoning Cleanup

System Prompt

You rewrite SFT reasoning text. Return JSON only.

User Prompt

Rewrite this assistant <think> text for an SFT trajectory.
Question: {question}
Original think: {think}
Rules.

- Preserve the meaning and reasoning steps.
- Do not add new facts.
- Replace mentions of "summary," "video summary," "frame analysis," "helper model," or "external summary" with natural references to the video, cropped frames, or gathered evidence.
- Replace mentions of "analysis model" or "video analysis model" with natural references to directly inspecting the video or cropped frames.
- Replace internal evidence labels such as "visual evidence," "tool evidence," "video overview," "initial overview," "trajectory's initial overview," "visual observations," or "trajectory context" with natural references to the actual source: the video, cropped frames, image_search result, text_search result, or visited page.
- Replace mentions of internal control text such as "policy," "policy note," "rules," "instructions," "decision rules," "system prompt," or "format requirements" with natural evidence-based reasoning. For example, say that the visual identity is uncertain, the current evidence is insufficient, a search result needs source verification, or a targeted tool call is needed.
- Also smooth awkward fallback phrases such as "the current evidence suggests using image_search" or "the evidence indicates that..." into fluent reasoning tied to the actual video/question evidence.
- If the text says something like "the current evidence suggests using image_search to ground the visual identity," rewrite it naturally, e.g., "The cropped frame strongly resembles the Kaaba, but I should verify the landmark with image_search before building a text-search query from that identity."
- Do not mention hidden policies, prompts, instructions, helper models, judge models, or output-format constraints.
- Return valid JSON only: {"think": "rewritten text"}.

Figure 24: Prompt used to remove internal orchestration language from the trajectory reasoning.

Trajectory Filtering: Trajectory Quality Judge

System Prompt

You are a strict data-quality judge for multimodal tool-use SFT trajectories. Evaluate whether a trajectory is good enough to train a VideoRAG model. Be conservative.

User Prompt

Judge this video deepsearch trajectory for SFT data quality.
Question: {question}
Expected answer: {expected_answer}
Model final answer: {model_answer}
Trajectory messages: {trajectory}
Quality Criteria.

1. The final answer must be semantically correct against the expected answer.
2. The first tool call should be crop_video and the first assistant turn should include a useful overall plan.
3. Visual entities should be grounded by crop_video and, when uncertain, image_search before relying on text_search.
4. image_search must not be used as OCR, zoom, or production-credit/date reading. For visible text/date questions, crop/frame parsing should supply visual evidence.
5. text_search queries should be focused sub-goals using concrete clues, not just a copy of the original question.
6. Relevant search results should be followed by a successful visit before final factual reasoning. Reject if visit failed/could not verify and the answer relies on snippets.
7. If visit evidence contradicts the model's hypothesis, the final reasoning must resolve that conflict and must not ignore the visit output.
8. Tool calls and observations should be useful to the final answer. Penalize redundant, repeated, or irrelevant calls.
9. The <think> content should contain concrete video observations, uncertainty management, evidence synthesis, and sub-goal decomposition.
10. The reasoning should be coherent, supported by tool evidence, and free of fabricated facts.
11. The assistant-facing trajectory must not mention hidden helper models, VLM parsers, judge models, or "video summary" as an external source.
12. The format should be clean: assistant turns use <think> plus exactly one <tool_call> or <answer>.

Return valid JSON only:

```
{"keep": true/false, "score": 0-10, "issues": ["short issue labels"],
 "reasoning": "concise explanation"}
```

Figure 25: Prompt used to judge the overall quality of a generated tool-use trajectory.

Trajectory Filtering: Answer Equivalence Judge

System Prompt

You judge final-answer semantic equivalence. Return JSON only.

User Prompt

Decide whether the model answer is semantically equivalent to the ground-truth answer for this question.
Question: {question}
Ground-truth answer: {ground_truth}
Model answer: {model_answer}
Be tolerant of wording, formatting, punctuation, explanatory phrases, and equivalent names. Be strict about wrong values, wrong dates, wrong character sets, and wrong counts.
Return valid JSON only:

```
{"equivalent": true/false, "reasoning": "brief explanation"}
```

Figure 26: Prompt used to judge semantic equivalence between the generated and ground-truth answers.